1-1-2013

Innovation Tournaments: Improving Ideas through Process Models

Joel O. Wooten
University of Pennsylvania, jwooten@wharton.upenn.edu

Innovation Tournaments: Improving Ideas through Process Models

Abstract

Innovation tournaments have a long history of driving progress, especially in the fields of engineering and design, and are once again gaining popularity thanks to advances in technology. Stripped to its essence, an innovation tournament is a process that uncovers exceptionally good opportunities by considering many raw opportunities at the outset and selecting the best to survive. Both the host of the tournament (the administrator) and the participants (the agents) face many decisions throughout this process. In the following papers, we answer a series of questions about innovation tournaments, addressing the specific managerial challenges of how to provide in-process feedback, how to moderate entry visibility, and how to understand and affect leaps in innovation. We report on two sets of field experiments using web-based platforms for graphic design contests and a unique data set from an online platform dedicated to data prediction tournaments. The answers to these questions contribute new understanding to the literature on innovation tournaments and offer managers guidance on improving outcomes.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Operations & Information Management

First Advisor

Karl T. Ulrich

Keywords

contest, creativity, idea generation, ideation, innovation tournament, open innovation

Subject Categories

Business

INNOVATION TOURNAMENTS: IMPROVING IDEAS THROUGH PROCESS MODELS

Joel Orba Wooten

A DISSERTATION

in

Operations and Information Management

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

____

Karl Ulrich, Professor, Operations and Information Management

Graduate Group Chairperson

____

Eric Bradlow, Vice Dean and Director, Wharton Doctoral Programs

Dissertation Committee

Karl Ulrich, Professor, Operations and Information Management

Christian Terwiesch, Professor, Operations and Information Management

Z. John Zhang, Professor, Marketing

ACKNOWLEDGMENT

These three papers were written alongside my advisor, Karl T. Ulrich, who first introduced me to innovation tournaments and made these efforts both possible and fun. The Mack Center for Technological Innovation at The Wharton School provided financial support for this research. Laura Kornish, Christian Terwiesch, and Z. John Zhang provided helpful feedback on portions of the manuscript. Participants in the 8th Annual Conference on Product and Service Innovation at the University of Utah improved portions of this paper with useful comments.

ABSTRACT

INNOVATION TOURNAMENTS: IMPROVING IDEAS THROUGH PROCESS MODELS

Joel O. Wooten

Karl Ulrich

Innovation tournaments have a long history of driving progress, especially in the fields of engineering and design, and are once again gaining popularity thanks to advances in technology. Stripped to its essence, an innovation tournament is a process that uncovers exceptionally good opportunities by considering many raw opportunities at the outset and selecting the best to survive. Both the host of the tournament (the administrator) and the participants (the agents) face many decisions throughout this process. In the following papers, we answer a series of questions about innovation tournaments, addressing the specific managerial challenges of how to provide in-process feedback, how to moderate entry visibility, and how to understand and affect leaps in innovation. We report on two sets of field experiments using web-based platforms for graphic design contests and a unique data set from an online platform dedicated to data prediction tournaments. The answers to these questions contribute new understanding to the literature on innovation tournaments and offer managers guidance on improving outcomes.

TABLE OF CONTENTS

LIST OF TABLES

Idea Generation and the Role of Feedback

The Impact of Visibility in Innovation Tournaments

Leaps in Innovation

LIST OF ILLUSTRATIONS

Idea Generation and the Role of Feedback: Evidence from Field Experiments with Innovation Tournaments

In many innovation settings, ideas are generated over time and managers face a decision about if and how to provide in-process feedback to the idea generators about the quality of submissions. In this paper, we use design contests allowing repeated entry to examine the effect of in-process feedback on idea generation. We report on a set of field experiments using two online contest websites to compare the performance of three different feedback treatments – *no* feedback, *random* feedback, and *directed* feedback (i.e., in-process feedback highly correlated with the final quality rating of the entry). We posted six logo design contests for consumer products and accepted submissions for one week. We provided daily feedback during the contest period using one of the three treatments. We then used a panel of target consumers to rate the quality of each idea. We find that directed feedback is associated positively with the quality of entries submitted and that quality improves with cumulative entries. In the aggregate, the variance in quality is not different across the three treatments. However, under directed feedback, the variance in quality declines as the contest progresses. We also find that the likelihood that an agent submits multiple entries increases in the presence of directed feedback.

1 Introduction

In many innovation settings, ideas are generated over time and managers face a decision about if and how to provide in-process feedback to the idea generators about the quality of the ideas submitted. On the one hand, such feedback could help avoid effort wasted exploring impoverished territory. On the other hand, feedback could over-determine the process, providing guidance that precludes exploration of unlikely but potentially very valuable directions.

In this paper, we use design contests allowing repeated entry to examine the effect of in-process feedback on idea generation. We report on a set of field experiments using two online contest websites for logo design to compare the performance of three different feedback treatments – no feedback, random feedback, and directed feedback (i.e., in-process feedback highly correlated with the final quality rating of the entry). In these experiments, we posted six logo design contests for consumer products and accepted submissions for one week. We provided daily feedback during the contest period using one of the three treatments. At the conclusion of the contests, we used a panel of target consumers to rate the quality of each idea. This approach is unique in its use of real competitions and designers in the experiments.

While feedback can be applied to most types of idea generation processes, we look specifically at *contests* or *open innovation tournaments*. Innovation tournaments have been used for high-profile innovation challenges, including Netflix's $1M Prize for improved movie recommendations and the X Prize Foundation's $10M price for private manned spaceflight. Despite a growing reliance on innovation tournaments in practice, relatively little research prescribes how to manage them more effectively. Our research aims to address a specific managerial challenge – if and how to provide feedback to contestants during the process of generating ideas. To answer this question, we look at three process parameters from the statistical view of innovation – average quality, variance in quality, and number of ideas – each of which influences the overall outcome.

We find that the type of feedback is indeed associated with differences in the idea generation process. Directed feedback is associated with higher average quality than no feedback or random feedback. We also find that quality improves with cumulative entries. Indeed, in the presence of directed feedback, quality "ratchets up" in association with the quality of the best prior entry by others and by the contestant. In the aggregate, the variance in quality of the ideas generated is not different across the three treatments. However, variance declines as the contest progresses under directed feedback. Finally, the likelihood that a contestant submits multiple entries increases in the presence of directed feedback.

2 Innovation Tournaments

Stripped to its essence, an innovation tournament is a process that uncovers exceptionally good opportunities by considering many raw opportunities at the outset and selecting the best to survive (Terwiesch and Ulrich 2009). Tournaments accomplish this through a series of steps that solicit and evaluate entries. Both the host of the tournament (the *administrator*) and the participants (the *agents*) face many decisions throughout this process. In the following section, we develop a conceptual model of tournaments and outline the decision variables for the administrator and agent. We believe this framework is valuable not only in setting up the specific research question we examine empirically, but—given emerging interest in innovation tournaments—as a review of the relevant scholarly research in innovation tournaments more generally.

2.1 A Conceptual Model

The administrator of an innovation tournament encounters a number of decisions, such as the length of the contest and how many rounds it will entail. Potential agents have their own choices to make, such as whether to enter and how much effort to invest. We used both bottom-up and top-down approaches to identify these decisions. In the bottom-up approach, we examined the literature on innovation tournaments and listed the decisions used in those papers, whether explicit or not. In the top-down approach, we considered the temporal flow of a tournament and what decisions would have to be made over the course of a contest. After combination and synthesis, nineteen distinct decisions emerged (Table 1). Although we believe these decisions are mutually exclusive and collectively exhaustive, inevitably the decisions reflect subjective judgments about groupings and the appropriate level of detail. For example, the *problem specification* decision represents several smaller decisions (e.g., problem breadth/scope, degree of specification, output required, contest platform). Other researchers would likely derive a similar but certainly not identical set of decisions.

Fundamentally, an innovation tournament works by defining a challenge, soliciting entries, moderating the contest, evaluating entries, and awarding a winner. In parallel, agents decide if and how to participate. Table 1 identifies nineteen decision variables for a tournament and divides them into six categories:

1. *Defining Challenge* – What does the contest (problem, specification, etc.) look like?
2. *Soliciting Entries* – Who can participate?
3. *Moderating* – What is the in-process feedback/information loop?
4. *Evaluating* – How are entries judged?
5. *Awarding* – What prize is at stake for the winner(s)?
6. *Participating* – How does an agent choose to engage (effort, strategy, etc.)?

The nineteen decision variables in six categories serve as a framework for organizing what we know about innovation tournaments from the prior literature.

2.2 Innovation Tournament Literature

In addition to codifying the innovation tournament decisions, Table 1 also maps the body of papers that specifically addresses those decisions in the context of innovation tournaments. A substantial set of papers, largely within the field of economics, relates theoretically to tournaments. Boudreau et al. (2011) give a nice summary of that related stream of economics

Table 1: Innovation Tournament Decision Variables and Literature

Administrator Decisions					Agent Decisions
Defining Challenge	**Soliciting Entries**	**Moderating**	**Evaluating**	**Awarding**	**Participating**
Problem specification Boudreau et al. (2011) Erat & Krishnan (2011) Kornish & Ulrich (2011)	*Participants* Bayus (2012) Poetz & Schrier (2012) Jeppesen & Lakhani (2010) Lakhani et al. (2007)	*Blind v. unblind* ---	*Judges* Kornish & Ulrich (2012) Soukhoroukova et al. (2007)*	*Type of prize/recognition* Terwiesch & Xu (2008)	*Enter v. not* Bockstedt et al. (2011) Walter & Back (2011)* Yang et al. (2008)
Duration of contest Yang et al. (2010)	*Number admitted* Boudreau et al. (2011) Terwiesch & Xu (2008) Che & Gale (2003)** Fullerton & McAfee (1999)** Taylor (1995)**	*Feedback* ---	*Evaluation criteria* Terwiesch & Ulrich (2009) Toubia & Flores (2007) Piller & Walcher (2006)	*Number of awards* Archak & Sundararajan (2010)* Morgan & Wang (2010)	*Effort invested* Bockstedt et al. (2011) Yang et al. (2008)
Number of rounds Gradstein & Konrad (1999)**	*Teams v. individuals* Girotra et al. (2010)		*Sychronous v. asynch.* ---	*Disclosure of winner(s)* ---	*Search strategy* Lakhani et al. (2007) Dahl & Moreau (2002)
Intellectual property rights Scotchmer (2004)**					*Collaboration with rivals* Bullinger et al. (2010)

Note: Many papers discuss multiple innovation tournaments decisions; we attempt to categorize the papers into the most representative bucket(s).
**Information Systems literature, **Economics literature*

literature on relative performance and incentives of contests in a variety of domains (i.e., Lazear and Rosen 1981, Holmstrom 1982, Casas-Arce and Martínez-Jerez 2009). Given that the economics literature has been summarized elsewhere, and that in most cases the links to innovation tournaments are somewhat tangential, we exclude those papers from the table.

2.3 Decision Variables

Contests have been shown to be effective platforms for innovation (Terwiesch and Xu 2008, Terwiesch and Ulrich 2009). The reduction in effort from any one participant that results from increased competition and negative incentive effects (Taylor 1995, Fullerton and McAfee 1999, Che and Gale 2003) can be offset by a larger participant pool and the positive effect of parallel exploration (Terwiesch and Xu 2008). Boudreau et al. (2011) empirically show these two effects to be of comparable magnitudes, with the biggest upside occurring in cases of problem uncertainty. Given that a tournament will be used, the first category of decisions facing an administrator is *Defining Challenge*. The prior literature shows that higher uncertainty problems mitigate the negative effect of lots of competitors (Boudreau et al. 2011), under-specifying a problem can be optimal (Erat and Krishnan 2011), and the size of solution spaces can be quantified (Kornish and Ulrich 2011). Yang et al. (2010) analyze broad contest characteristics from TaskCN.com and determine that lengthier project specifications delay submissions – but attract the same number of agents – and that doubling project duration increases participation by 34%. Preliminary work has also looked at the considerations for multiple rounds (Gradstein and Konrad 1999) and patents in prize contests (Scotchmer 2004).

An administrator's decisions within the *Soliciting Entries* category influence the types of solvers and specific individuals who participate. In a study of 166 science contests from InnoCentive.com, decisions around breadth of solicitation and level of expertise revealed the value of openness and broadcasted search to include non-obvious solvers (Jeppesen and Lakhani

2010)[1]. Girotra et al. (2010) examine group structures and find that individual-hybrid groups are better able to generate and filter ideas than pure teams.

When *moderating* an innovation tournament, the administrator faces the decision of whether or not to allow participants to see the submissions of their rivals (i.e., *blind v. unblind*). To our knowledge there has not been prior published research on this question. The administrator also faces a decision on the nature of the in-process feedback provided to the agents in response to their submissions. This decision is the focus of our empirical study, and we discuss the related literature in greater detail in Section 3.

Within the *Evaluating* category, we know that surveys of consumers appear more reliable than expert evaluations (Kornish and Ulrich 2012); Soukhoroukova et al. (2007) propose using idea markets as an alternative to either experts or participants. In terms of the criteria used, the best performing schemes are those that focus on potentially misclassified ideas and avoid dismissing ideas too quickly (Toubia and Flores 2007). Kornish and Ulrich (2012) further characterize a method for evaluating the performance of an idea selection process.

In terms of *Awarding*, Terwiesch and Xu (2008) demonstrate that performance-contingent awards can offset agent underinvestment better than fixed-price awards. With risk-neutral agents, administrators should allocate one prize; with risk-averse agents, multiple prizes can be optimal (Archak and Sundararajan 2009).

In contrast to the decisions that a tournament host faces, agents each face certain *Participating* decisions. Natural experiments on LogoMyWay.com show that earlier entrants to unblind contests are more likely to win, as are those with a wider range of entry timing; however, simply increasing the number of entries does not benefit the agent (Bockstedt et al. 2011). Walter and Back (2011) and Yang et al. (2008) collect data from other online markets and show varying

[1] Lakhani et al. (2007), Bayus (2011), and Poetz and Schrier (2012) complement this view, showing the value of varied interests, new serial ideators, and end users (vs. professionals), respectively.

levels of agreement in relating contest measures to agent behavior. With regard to how agents search for successful solutions, analogical thinking (Dahl and Moreau 2002), recombination of previously-developed expertise (Lakhani et al. 2007), and cooperation among competitors (Bullinger et al. 2010) can be effective. In one study of 166 scientific problems posted as innovation contests on the Innocentive website, intrinsic motivation was the leading driver of performance in the winners (Lakhani et al. 2007).

3 The Role of Feedback

In-process feedback is a critical component of innovation processes (Kline and Rosenberg 1986). It introduces information signals that steer future development. However, there has been no prior work focused on feedback in innovation tournaments. One study does raise several questions around tournament feedback and offers an anecdotal perspective (N=1) on its importance (Yang et al. 2010) but leaves the questions unanswered. Despite this dearth, there is some research on feedback in problem solving and contest settings more generally. We turn to this work to help develop hypotheses about the role of feedback in the innovation tournament setting.

3.1 Feedback Literature

The type of contest feedback most often studied involves revealing participant skill level in head-to-head competitions for well-defined tasks (e.g., solving mazes or math problems). In particular, lab experiments show that while top-performing competitors do well with feedback, performance – but not effort – deteriorates for the worst-performing competitors as feedback frequency increases (Bull et al. 1987, Eriksson et al. 2008). In a similar study, Freeman and Gelber (2009) demonstrate that a greater possibility of a prize dramatically increases the performance of the bottom half of competitors with full information. In an observational study of UK students, unconfident individuals getting a good score produced more effort, whereas overconfident students getting a good score slacked (Bandiera et al. 2009). These suggest that individuals getting unexpected, good feedback can increase both effort and performance in certain contests.

The field of behavioral psychology, while not associated specifically with innovation, has long studied such responses to stimuli and methods of modifying behavior. In a famous early example, Skinner (1948) ran a study on pigeons' responses to rewards in the form of food; the birds think they are being rewarded for particular actions, so they continue performing and reinforcing random behaviors (e.g., a head-turn to the left). The feedback in this setting, which resembles random feedback in our study, reinforces behavior, whether good or bad. More recently, a meta-analysis covering over 130 psychology studies revealed a heterogeneous effect of feedback on performance, with overall improved performance but decreased performance in over one-third of treatments (Kluger and Denisi 1996).

This context-dependent nature of feedback is reinforced by a crowd-sourced survey experiment from labor economics that finds feedback lowers rates of task reentry and lowers productivity, except in top performers (Barankay 2011). The negative relationship between feedback and effort differs from some of the previously mentioned competition contests and could have negative effects in an innovation tournament.

The management literature addresses the role of feedback and communication in organizational dynamics. Tjosvold and McNelly (1988) demonstrate that the quality and type of communication, rather than its frequency, improve organizational innovation – a finding that supports the theory that interaction, feedback, and access to more information will lead to greater levels of innovation. Several empirical studies link higher levels of information gathering and both internal and external group communication with better performance in research and development groups (e.g., Katz 1982, Keller 1986).

Communication and idea generation have also been addressed in the social psychology literature. Highsmith (1978) posits that a lack of meaningful, positive feedback greatly reduces the rate of idea generation in group sessions. In a simulated study around organizational context, the average number of ideas combined goes down in the absence of communication. If there is variability in communication, however, there is no effect on the average number of proposals combined or on the variance in their quality (Seshadri and Shapira 2003).

In sum, the literature presents us conflicting theories about which effects might dominate in innovation tournaments. On one hand, the presence and accuracy of feedback could improve the rate and quality of idea generation, hone the search for a solution, and help avoid effort wasted exploring impoverished territory. On the other hand, the presence and accuracy of feedback could over-determine the process, providing guidance that precludes exploration of unlikely but potentially very valuable directions. To our knowledge, there have been no empirical explorations of these effects in innovation tournaments.

3.2 Defining Feedback

We focus on the specific innovation tournament setting in which individual agents may submit ideas repeatedly to an open, unblind, moderated contest. The tournament is *open* in the sense that anyone may choose to participate. The tournament is *unblind* in the sense that all ideas and feedback are visible to all participants. The tournament is *moderated* in the sense that the contest administrator may provide feedback on the quality of submissions. Recall that we consider three different types of feedback:

1. Directed feedback. Information provided to the agent on the quality of an idea, shortly after its submission, is highly correlated with the administrator's quality function, which is the final determinant of performance.
2. Random feedback. Information provided to the agent on the quality of an idea, shortly after its submission, is largely uncorrelated with the administrator's quality function.

3. No feedback. No information on the quality of an idea is provided to participants.

In most innovation tournaments, overall performance is determined by the quality of the best ideas – the quality of the ideas in the upper tail of the distribution. Indeed, in open tournaments, rewards are typically only awarded to the best ideas.

The notion of best is defined by the quality function of the administrator. This function is rarely an explicit mathematical expression. Sometimes, as with the X-Prize tournaments, testing of various kinds is used to determine quality. More typically it is a subjective judgment by one or more evaluators. In most cases, the quality function of the administrator is not arbitrary, but rather is similar to those of agents knowledgeable about the domain of the tournament. For instance, in logo design, participating graphic designers are likely to agree to some extent on what is good design. Despite some shared understanding, the quality function of the administrator is rarely, if ever, known perfectly.

3.3 Measuring Impact

In order to evaluate the ideas generated in our tournaments, we rely on the statistical view of innovation. This perspective, in which idea creation is a series of random draws from a distribution, was developed by Dahan and Mendelson (2001) and then further elaborated by Girotra et al. (2010). A key insight from this literature is that the success of idea generation in innovation depends not on the entire body of opportunities identified, but rather on the quality of the *best* ideas; the extremes are important, not the average or the norm (March 1991; Dahan and Mendelson 2001; Terwiesch and Ulrich 2009; Girotra, et al. 2010). Using this framework, Girotra et al. (2010) showed that the best idea from an idea generation process depends on three process parameters: the average quality of ideas generated, the variance in the quality of ideas generated, and the number of ideas generated. Since tournament performance is dictated by extreme values, the actual quality of the winning outcomes is very noisy, and so the quality of the winning submission is not a practical dependent variable in an empirical study. Following in the tradition of the prior literature, we focus instead on the three process parameters that drive the outcome.

3.4 Impact of Feedback on the Quality of Ideas Submitted

When generating ideas for a moderated tournament allowing repeated entry, individuals engage in independent, parallel exploration. Each agent has his or her own imperfect understanding of the administrator's quality function. Agents explore their own landscape of possibilities, decide which ideas are most promising, develop those ideas, and submit one or more entries. The contest itself provides two sources of ongoing information to participants:

1. The ideas submitted by others, which may illuminate the landscape of possibilities, even absent feedback from the administrator.
2. Feedback from the administrator on the ideas of both the agent and others. The agent is likely to use this feedback to update his or her understanding of the quality function of the administrator.

Making use of this new information, agents may choose to engage in additional exploration and to submit additional entries.

The fundamental idea that information can lead to learning is explored in the literature on mental models. Mental models activate when new information is incorporated into one's base of knowledge, resulting in conceptual change. Enrichment occurs in the simple case when consistent information reinforces the existing framework, and revision happens when the new information is inconsistent with prior beliefs (Vosniadou 1994). Vosniadou goes on to point out that learning failures are more likely when revisions are needed, which can produce inconsistencies. This suggests that feedback schemes that increase the amount of accurate, accretive information will reduce misconceptions, enhance learning related to the quality function, and thereby improve the average quality of submissions.

Based on this logic, we would expect agents to learn the most, and therefore produce better ideas on average, when the administrator provides directed feedback. Furthermore, we would expect learning failures and relatively lower performance in the face of random feedback. Under the no-feedback condition, the agents still have new information based on the submissions of others, which we would expect to enhance learning and therefore average quality. However, we would not expect agents under this treatment to perform as well on average as with directed feedback.

Because learning is likely to occur both from information revealed by the entries of others and from the feedback from the administrator, we also expect that the quality of ideas submitted by an agent will increase as the agent observes more cumulative entries. Furthermore, we expect quality to increase as ideas of better quality are revealed, a sort of ratcheting up of quality. So, an idea of high quality submitted by one agent, by revealing a promising direction for exploration, is likely to increase the quality of subsequent entries. This ratcheting up in quality is also likely to play out for individual agents. An individual agent is unlikely to submit ideas that are clearly inferior to those he or she has submitted previously. We can pose these expectations as the following testable hypotheses.

> *Hypothesis 1*: The quality of ideas submitted is increasing in the accuracy of the feedback provided, with directed feedback the most accurate, no feedback the next most accurate, and random feedback the least accurate.
>
> *Hypothesis 2*: The quality of ideas submitted is increasing in the number of cumulative contest entries.
>
> *Hypothesis 3*: The quality of ideas submitted is increasing in the quality of the best previous idea submitted by anyone.
>
> *Hypothesis 4*: The quality of ideas submitted by an agent is increasing in the quality of that agent's best previous entry.

3.5 Impact of Feedback on Variance in Quality of Ideas Submitted.

Variance in the quality of ideas submitted could be caused by variance in the skills and capabilities of the agents (i.e., across-agent variation) and/or by variance in the quality of multiple ideas submitted by a particular agent (i.e., within-agent variation). Differences in ideas may be due to differences in the basic approach taken and/or due to differences in the elaboration and execution of that approach. (In the logo design setting, the same raw idea may be manifest in ways that vary highly in quality.) We expect that in any open tournament, directed feedback may lead to a convergence in approaches, as agents seek to imitate the approaches that have worked for others, which is likely to reduce variance in quality. In the presence of random feedback, we might expect greater variation in the approaches attempted by agents in an attempt to discover the administrator's quality function.

We expect that as individuals construct a more accurate model of quality, variance in quality across their ideas will decrease. Thus, we would expect less variance in quality under the directed feedback condition as the number of entries in the contest increases. Whereas under the random feedback condition, the agent does not know the quality of the submissions and may act on inaccurate information about quality and so may explore otherwise unlikely directions. These expectations can be posed as these two hypotheses.

> *Hypothesis 5:* The variance in quality of ideas submitted is lowest under directed feedback and highest under random feedback.
>
> *Hypothesis 6*: The variance in quality of ideas submitted decreases under directed feedback as the number of cumulative entries increases.

3.6 Impact of Feedback on Number of Ideas Submitted

Once an agent has submitted an idea, he or she has the option to submit additional ideas. Two forces are likely to influence the likelihood of additional submissions. First, engagement with the administrator is likely to result in additional submissions. As discussed in Section 2, the social psychology literature, specifically studies by Highsmith (1978) and Sheshadri and Shapira (2003), documents increased rates of idea combination and generation in association with communication and feedback. This finding supports the hypothesis that agents will be less likely to engage in a tournament with repeated entry under the no feedback treatment. A lack of feedback may also diminish the sense of community, which seems to be an underlying motivation in many open innovation settings (Lakhani et al. 2007).

Second, an agent who receives feedback consistent with his or her own understanding of the quality function is likely to submit additional entries because the agent can be more confident in his or her own judgments and therefore in the likelihood of selecting ideas for submission that will be perceived as good. Directed feedback is more likely to result in a perception of the administrator "getting it" than is random feedback. In the case of random feedback, the agent may perceive that efforts are wasted and therefore lose motivation to submit.

Integrating these two forces, we would expect the directed feedback treatment to result in the highest likelihood of repeat submissions. We cannot predict whether random feedback or no feedback would result in lower rates of submission, as the two forces would act in opposite directions for these treatments, which leads to this hypothesis.

> *Hypothesis 7*: The incidence of repeat submissions by an individual agent is higher under directed feedback than under no feedback or random feedback.

4 Experimental Design

We conducted a set of six field experiments comparing the performance of three distinct treatments – no feedback, random feedback, and directed feedback. We posted three pairs of logo design competitions with substantially similar problem statements using two online marketplaces built around design contests, 99Designs and CrowdSpring. After the contests were completed, we used a consumer panel of judges to rate the quality of each entry.

4.1 Platforms

Several online companies have emerged as leaders in the crowd-sourced design market, which allows buyers to solicit projects – such as logo generation – from a community of artists and graphic designers. The administrator creates a contest by posting project specifications and a prize amount, and then receives online submissions from agents (graphic designers in this case). During a contest, the websites permit the administrator to provide several types of feedback to agents. Feedback can be posted in the form of public comments, private messages, ratings (1 to 5 stars), and entry elimination. After a prescribed time period, the administrator selects a winner and awards the prize. These websites have proven to be an inexpensive and popular way to gain access to a wide array of creative talent. The two websites used in this experiment were 99Designs and CrowdSpring (Table 2). They are very similar in their implementation, with well-designed and nearly identical user interfaces.

Table 2: Contest Website Comparison

	99Designs.com	CrowdSpring.com
Designers on website	38,658	27,000+
Average entries per contest	86	77
Minimum contest award	$150	$200
Active logo contests	215	76

Note: Website statistics one month prior to experiment

4.2 Contests

A total of three pairs of contests were created as follows.

Crazy Comet Soccer Gear	Mexicali BBQ Sauce	Bright Bay Toys
Supernova Swim Wear	South of the Border Salsa	Color Cove Games

One of each pair was randomly assigned to each of the two websites. All six contests had nearly identical details, including company type, name, design specifications, deliverables, target markets, and brief specifics. Each logo contest was for a new consumer products brand whose target audience was college-educated U.S. consumers 18-35 years old. The contests in each pair shared the type of product (sports gear, condiments, and toys/games). An example of the submitted design briefs is in Appendix A.

4.3 Treatments

The independent variable tested was the degree of feedback that accompanied each design contest. Feedback was delivered in the form of ratings once per day using a scale of one to five stars. The star mechanism is well established on these platforms as the signal for how much an administrator likes a design. For those receiving directed feedback, a panel of independent *feedback judges* determined the rating. For those that received random feedback, the rating was determined by a random number generator according to the probabilities listed in Appendix B. In the contests receiving no feedback, none of the entries received any ratings. Both websites algorithmically monitor administrator activity and flag inactive contests. Feedback factors heavily in this monitoring, so every agent was left a private thank you message, regardless of the treatment received, in order to ensure that all contests remained in good standing.

In order to provide star ratings for the two contests that received directed feedback, we recruited six individuals from the target market to act as feedback judges. All feedback judges were informed that this was an experiment involving logo creation contests and that participation was voluntary. The overall charge to feedback judges was aligned with the objective given to the designers – to develop a logo that will be most appealing to college-educated U.S. customers, age 18-35, for the given product. The feedback judges viewed logos daily in a randomized order and rated them on a 1-to-5 scale in response to the question, *To what extent does this logo appeal to you?* The feedback provided under the directed treatment was the average rating expressed as a number of stars.

4.4 Experiment

We denote the three pairs corresponding to the three product types as A, B, and C. One of each pair ran on 99Designs and the nearly identical corollaries ran on CrowdSpring, allowing for each of the three levels of feedback to be tested on each site and against each other in a balanced, incomplete block design. Designers closely monitor the contests on these websites and frequently report copyright violations and other such concerns. To deal with such savvy agents and avoid undermining the outcomes, we carefully constructed the experiment design to utilize two different website platforms and slightly staggered start dates for the three product sets. The feedback treatment can be denoted by subscripts (*N* for none, *D* for directed, and *R* for random) resulting in the following contest layout:

99Designs:	A_N	B_D	C_R
CrowdSpring:	$\overline{A}_D$	$\overline{B}_R$	$\overline{C}_N$

The six contests relied on the standard mechanisms of the websites to entice designers to participate. Each contest was open to anyone on the respective website, ran for seven days, received daily feedback (if applicable), and resulted in an award to the winning designer of $200. All contest entries were visible to the public, so agents could see every submitted logo and any rating feedback that was received. All experiments were conducted after obtaining approval from the human subjects committee at the university. An overview of the experiment is illustrated in Appendix C.

4.5 Evaluation

At the conclusion of the submission phases of the six contests, a total of 624 entries were generated by 245 designers. Some submissions included multiple logos, while others were duplicates or near duplicates. An eventual set of 544, whose progression can be seen in Appendix D, were independently and anonymously evaluated by a panel of 36 judges using a web-based survey in order to obtain a measure of the quality of each idea. The order of logos in the survey was randomized for each judge. The 36 judges were representative of the target market for the six contests – college-educated individuals in the U.S. between the ages of 18 and 35 – and they evaluated the logos from the perspective of potential consumers. These judges were similar but distinct from the feedback judges, who provided the daily feedback for the star ratings. All judges were informed that this was an experiment involving logo creation contests and that participation was voluntary.

Ideas from the six contests were rated in separate surveys and were completed in groups of two, corresponding to the product type (e.g., sports gear, condiments, toys/games). To mitigate order effects, surveys were administered as a balanced, repeated Latin square design in which the different sequences of surveys (ABC, ACB, BAC, etc.) resulted in each sequence appearing the same number of times. The surveys were sent in three batches. Each batch contained the matched pair of contests. Within each pair, half the judges were given the contest from 99Designs first followed by the one from CrowdSpring; the other half saw them in the opposite order. Within each individual survey, the logos were presented to each judge in a randomized order. The question and response choices (on a 1-5 rating scale) were the same for the judges as for the feedback panel. A threshold of 94 was established as the maximum number of logos to be graded in any one survey in order to limit surveys to a manageable ten minutes. This cutoff was established after survey pre-tests on a separate, but similar, group showed deteriorating results in longer surveys. Only one contest (Bright Bay Toys, with 188 ratable designs) exceeded this threshold. As a result, each judge rated all of the designs in the other five contests and exactly half of the designs in the Bright Bay contest. For the Bright Bay surveys, the assignment of particular logos to judges followed a balanced, repeated Latin square design in which every rater saw 94 logos and every logo received 18 raters.

The judges' responses to the six surveys provide the raw data for our analysis. The inter-rater reliability of judges is high. We check this using a Krippendorff alpha test for inter-rater reliability across the 36 judges. Given the artistic nature of our contests, personal preferences will yield variation in the judges' scores, and we observe this with a relatively low degree of agreement between any two judges (Appendix B). However, if populations have stable preferences, then a high degree of agreement should be seen in the average quality for each logo across two sample populations. We compare the average logo scores from our sample of 36 judges to the averages from the survey pre-test of 24 similar judges and obtain a degree of agreement of 0.74, well above commonly accepted thresholds. A sample of the scored logos is provided in Figure 1.

Figure 1: Illustrative Examples of Logos Generated

99: 99Designs, CS: CrowdSpring; N: no feedback, D: directed feedback, R: random feedback

5 Analysis and Results

Table 3 provides the summary statistics for the outcomes of each contest. Ideally, we would want to analyze performance at the level of the overall contest. Unfortunately, because contests are relatively expensive and logistically challenging, it is not feasible to run a large sample of distinct contests. Furthermore, the quality of winning submissions will also vary widely because the winner is the extreme value from a set of candidates – a statistic that exhibits a great deal of variation. We have just six observations at the level of the contest, and so any differences in outcomes at the contest level are not likely to be statistically significant. The six contests also vary in terms of the website they were run on and the types of products, which introduces confounding effects that cannot be controlled for in a sample of six. Instead we analyze the data at the level of the *entries* to the contests, of which there are hundreds, and at the level of judges'

Table 3: Contest Summary Statistics

	Sports Gear		**Condiments**		**Toys/Games**	
	Crazy Comet Soccer Gear	Supernova Swimwear	Mexicali BBQ Sauce	South of the Border Salsa	Bright Bay Toys	Color Cove Games
Feedback	None	Directed	Directed	Random	Random	None
Website	99Designs	CrowdSpring	99Designs	CrowdSpring	99Designs	CrowdSpring
N Logos	55	86	88	42	188	85
N Designers	30	38	35	27	70	44
Best Logo	3.44	3.50	3.36	3.61	3.60	3.64
2nd Logo	3.19	3.31	3.33	3.31	3.55	3.33
3rd Logo	3.06	3.28	3.31	3.22	3.49	3.22
4th Logo	3.03	3.22	3.28	2.94	3.45	3.19
5th Logo	2.86	3.14	3.25	2.86	3.37	3.19
Top 5 Mean	3.12	3.29	3.31	3.19	3.49	3.32
Overall Mean	2.37	2.47	2.47	2.41	2.33	2.47
S.D.	0.43	0.48	0.49	0.42	0.52	0.49

Note: Values listed are averages of judges' ratings on 1-5 scale; Bright Bay scores adjusted for judge.

ratings of entries, of which there are thousands. The bulk of our analysis examines differences in the quality, variance, and frequency of individual submissions to the contests with respect to the feedback treatment and other explanatory variables. In the discussion, we will relate these process-level results to the implications for overall contest performance.

More specifically, the dependent variable on which most of our analyses are based is the rating on a 1-to-5 scale of a particular logo by a particular judge. There are 16,200 such logo-ratings. To test our hypotheses, we regress the logo rating against the variables of interest (e.g., feedback treatment) and a set of control variables.

We include controls for the website on which the contest was conducted (the site), the paired product type (the domain), the time (both the contest day[2] and the day of the week), and the judge. This is to account for any differences in talent across platforms, any differences imposed by innate product characteristics, and any temporal differences resulting from the day. The control for the judge captures differences in how judges use the rating scale.

In all cases, the results tables show analyses run for one base case (i.e., omitted variable) – the directed feedback treatment dummy variable. Pair-wise comparisons of the coefficients for the non-omitted treatments are left for the discussion. Unless noted, variables are not centered, to permit easier interpretation of the contest progression from the start of the tournament.

5.1 Effect of Feedback on Quality of Ideas Submitted

Table 4 shows the results of a regression analysis with logo rating as the dependent variable and these independent variables: treatment (none, directed, or random), the agent's best prior feedback score (MaxStarI), cumulative number of entries, best prior entry by others (Max), and the agent's best prior entry (MaxI). Because there are multiple ratings for a single logo we use a

[2] *Contest days* are the 24-hour periods corresponding to the seven days that each experiment ran. So contest day 1 is the first 24 hours that the contest was live, regardless of the actual day of the week.

clustered OLS for the analysis when using ratings as the dependent variable.[3]

Before analyzing our hypotheses, the first question one might ask is whether the feedback given actually impacts the quality of a designer's next submission. Model 1 shows that the quality of first entries (those submitted before getting personal feedback) for all participating designers doesn't differ across treatments (N=8,072). This suggests that the populations arriving to these contests are similar and that feedback treatments do not affect the first submissions. Model 2 then shows that the judged quality of designers' second submissions differ, with more stars on first entries leading to higher-quality second entries under directed feedback and lower-quality second entries under random feedback.

If our contest setting revolved around a single entrant, then the preceding feedback-quality improvement link would be sufficient. However, in open, unblinded contests with repeated entry, a greater level of analysis is needed. Imagine a case where every entrants' second design was worse than their first, but each entrants' first design was better than the first design that came before it. That would result in a better tournament outcome, despite a failing in the individual feedback response. So while the answer to the previous question is interesting, it is not necessarily sufficient in the context of contests.

To test our hypotheses, we return to idea generation as a random draw and test the differences in quality for distinct submissions. Models 3 and 4 show the analysis for all logo-ratings (N=16,200). Model 5 shows the analysis only for those logo-ratings associated with a repeated entry by an agent (N=8711).

When including as explanatory variables only the feedback treatment and the control variables, directed feedback and no feedback are associated positively and significantly with the quality of a submission relative to random feedback. (There is no significant difference between

[3] If we did not use a cluster correction, our OLS standard error estimates could be biased downward as a result of explanatory variables varying at one level (that of the contest or logo) and our responses being sampled at the individual rating level. For robustness, we examine additional models on subsequent pages.

Table 4: Comparison of Quality between Feedback Treatments

	Ratings of 1st Entries	Ratings of 2nd Entries	All Ratings	All Ratings	Ratings of Repeaters
	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*
Intercept	2.887 ***	2.768 ***	2.645 ***	2.683 ***	0.341 *
	(0.104)	(0.234)	(0.099)	(0.217)	(0.207)
Control Variables					
see notes	⊥	±	†	‡	Ŧ
Treatment					
None	-0.102	0.476	-0.057	0.550 **	0.925 ***
	(0.090)	(0.393)	(0.061)	(0.259)	(0.276)
Random	-0.046	-0.019	-0.162 **	0.346	0.573 **
	(0.096)	(0.157)	(0.064)	(0.321)	(0.264)
Effects					
MaxStarI		0.193 **			
		(0.097)			
MaxStarI x Trt[None]		±			
MaxStarI x Trt[Random]		-0.280 **			
		(0.109)			
Entries				0.003	
				(0.002)	
Entries x Trt[None]				0.007 **	
				(0.004)	
Entries x Trt[Random]				0.000	
				(0.002)	
Max				-0.001	
				(0.073)	
Max x Trt[None]				-0.241 **	
				(0.097)	
Max x Trt[Random]				-0.134	
				(0.118)	
MaxI				0.026	0.874 ***
				(0.027)	(0.066)
MaxI x Trt[None]				-0.002	-0.368 ***
				(0.041)	(0.110)
MaxI x Trt[Random]				-0.029	-0.250 **
				(0.039)	(0.106)
R-squared	0.13	0.21	0.13	0.13	0.22
Mean Response	2.45	2.41	2.42	2.42	2.40
Observations	8,072	1,319	16,200	16,200	8,711
DF	40	42	47	56	50

OLS regression on individual ratings, clustered by logo, base case is Directed feedback
Significance levels: * <0.10, ** <0.05, *** < 0.01
Robust clustered standard errors given in parentheses

⊥ Not sig.: Domain, Site — Sig.: Judge — Robust to: 1st single entry, Time controls

± Not sig.: Domain, Site — Sig.: Judge — Robust to: Time controls
Ratings of 1st and 2nd entries (DVs) group entries from that day
Interaction omitted (No feedback has no MaxStarI levels); MaxStarI mean-centered around 2.82

† Not sig.: Domain, Site, Day — Sig.: Judge, Contest Day — Robust to: Ordered probit model, # agents
‡ Not sig.: Domain, Contest Day — Sig.: Judge, Site, Day — Robust to: Ordered probit model
Ŧ Not sig.: Contest Day, Day — Sig.: Judge, Domain, Site

directed feedback and no feedback.) These results support Hypothesis 1 to the extent that random feedback is associated with lower quality than either directed feedback or no feedback, although they do not allow us to distinguish between no feedback and directed feedback.

When including the cumulative entries submitted and the quality of the best of those entries submitted, a more nuanced association emerges, which is supportive of Hypotheses 2 through 4.

Under the no-feedback condition, the cumulative number of prior entries has a significant positive association with the quality of a submission. In the presence of feedback (whether random or directed), the magnitude of this association is significantly lower. An interpretation of these results is that when there is no feedback, the agent focuses on learning from the prior submissions to the contest. However, when there is feedback, whether or not it is accurate, the agent seeks information from the feedback ratings, and thus pays less attention to the prior entries.

In contests with directed feedback, the best prior entry from anyone (*Max*) is associated positively and significantly with the quality of a subsequent entry. One interpretation of this result is that if feedback is available and accurate, then the quality of prior entries serves as useful information in improving subsequent submissions. However, if the feedback is random or nonexistent, then agents are not able even to identify the best prior entries, and therefore cannot use them effectively as signals.

The last rows of coefficients in Table 4 reflect the role of the agent's own best prior entry (*MaxI*) on the quality of subsequent entries. In the full data set, no effect is significant. However, the majority of designers submit just one logo. For this reason, we conduct the analysis (Model 5) for repeat entries only. We find a significant, positive association between an individual's best prior entry and their future entries under all three treatments. Of course the main effect can be interpreted as the skill level of the agent – agents producing one good design are likely to produce others. However, when considering the interaction with the type of feedback, there is a significantly higher association between the quality of the agent's best prior submission and subsequent entries under directed feedback than under no feedback or random feedback.

The above results are robust to several approaches with respect to time controls and regression model. Using just a designer's first entry in Model 1, instead of averaging all entries submitted on the first day (before receiving feedback), produces similar results. Models 1 and 2 can introduce time controls, such as contest day and day of the week, although both variables lose some of their meaning when averaged across entries and were omitted in Table 4. To check robustness around the analysis of our dependent variable, we use an ordinal probit model and obtain the same results as shown in Model 3 and Model 4.[4] An ordinal model is useful if the distances between the ratings levels are not equidistant, which does not appear to be the case with our ratings data.

5.2 Effect of Feedback on Variance in Quality of Ideas Submitted

Table 5 reports the results of an analysis in which the dependent variable is the squared difference between the average quality rating of an entry and that of the average ratings of all logos at that time in the contest (N=544).[5] This is a measure of variance that accounts for a linear quality improvement trend over time, which is a prominent feature of these contests.

We again look first at the impact of feedback on an individual's second submission. Model 6 shows that the variance of a second entry (those submitted after getting personal feedback) differs, with more stars on first entries leading to higher-variance second entries under random feedback and lower-variance second entries under directed feedback. Having some understanding

[4] While ordered probit models are particularly well suited to ordinal dependent variables, they become unstable if cells are empty or small. In this case, the crosstab of our categorical and response variables show some empty cells (judge and rating pairs). In addition, the responses, while ordinal in measurement, represent a continuous concept (quality) and there is indication that intervals between the points are approximately equal. For these reasons, we present the OLS model as the primary analysis, whose coefficients are also easier to interpret.

[5] We take the squared difference variance calculation from Girotra et al. (2010) and modify it slightly to account for the fact that our contests run for one week and thus have a temporal effect.

Table 5: Comparison of Variation between Feedback Treatments

	Variances of 2nd Entries		All Variances		All Variances	
	Model 6		*Model 7*		*Model 8*	
Intercept	0.260		0.169	***	0.249	***
	(0.183)		(0.055)		(0.068)	
Treatment						
None	-0.455		-0.057		-0.184	**
	(0.309)		(0.042)		(0.080)	
Random	-0.127		-0.030		-0.187	**
	(0.163)		(0.045)		(0.085)	
Control Variables						
Domain[Condiments]	-0.165		-0.002		0.017	
	-0.133		(0.041)		(0.044)	
Domain[Toys/Games]	-0.028		0.052		0.069	
	(0.165)		(0.045)		(0.051)	
Site[CrowdSpring]	-0.112		-0.015		-0.019	
	(0.112)		(0.030)		(0.042)	
Contest Day	0.026		0.004		0.013	
	-0.023		(0.007)		(0.019)	
Day	Ŧ		‡		†	
Effects						
MaxStarI	-0.106					
	(0.083)					
MaxStarI x Trt[None]	Ŧ					
MaxStarI x Trt[Random]	0.192	**				
	(0.094)					
Entries					-0.003	*
					(0.002)	
Entries x Trt[None]					0.003	*
					(0.002)	
Entries x Trt[Random]					0.003	**
					(0.001)	
R-squared	0.205		0.03		0.04	
Mean Response	0.18		0.23		0.23	
Observations	47		544		544	
DF	9		12		15	

OLS regression on quality de-trended variation at logo level, adj. for judge, base case is Directed feedback
Significance levels: * <0.10, ** <0.05, *** < 0.01
Standard errors given in parentheses

‡ Omitted control variables, significant
† Omitted control variables, not significant

Robust to: Time periods instead of entries as measure of contest progression; Quantile regression of the effect on feedback on the individual, we now test our hypotheses over the entire contest.

In the aggregate (Model 7), there are not significant differences in variance across the three treatments, which is not consistent with Hypothesis 5. However, there are significant trends in the level of variance as a contest progresses. When including number of prior entries as an explanatory variable (Model 8), there is a decline in variance over time under the directed feedback treatment. The net effect is that variation under directed feedback is initially higher than under the other treatments, but becomes lower after about 62 entries. These findings are consistent with Hypothesis 6. An interpretation of these results is that as entries are submitted under directed feedback, a consensus emerges as to the direction of the best designs. As this consensus develops, there is less and less variation in the quality of submissions.

5.3 Effect of Feedback on Number of Ideas Submitted

We hypothesized (H7) that the feedback treatment influences the likelihood that an agent submits repeatedly to a contest. Model 9 in Table 6 reports on a probit analysis including control variables and model effects, as well as cohorts, which control for multiple entries from an agent with no opportunity for feedback. We observe that directed feedback is associated significantly and positively with agents submitting multiple entries. We observe from this analysis that both random and no feedback are associated significantly and negatively with agents submitting multiple entries. This result supports Hypothesis 7. An interpretation of this result is that designers who receive feedback consistent with reasonable expectations of quality are more likely to continue to engage in the competition. This feedback may give the agent confidence that he or she understands the contest and that the administrator is engaged.

We corroborate this finding with a negative binomial regression (Model 10), which is a generalized Poisson that can accommodate over-dispersed data. The dependent variable is a count of the number of submissions by each designer.

Table 6: Comparison of Productivity between Feedback Treatments

	Probit Regression: Probability of Repeater		Negative Binomial Regression: # of Logos per Entrant	
	Model 9		*Model 10*	
Intercept	-0.748	*	1.019	***
	(0.393)		(0.203)	
Treatment				
None	-1.108	**	-0.807	***
	(0.479)		(0.284)	
Random	-0.965	**	-0.766	**
	(0.492)		(0.297)	
Control Variables				
Domain[Condiments]	-0.106		-0.057	
	(0.257)		(0.155)	
Domain[Toys/Games]	0.299		0.267	
	(0.314)		(0.183)	
Site[CrowdSpring]	-0.526	*	-0.294	*
	(0.265)		(0.151)	
Contest Day	0.180		-0.096	
	(0.116)		(0.062)	
Day	‡			
Effects				
Entries	-0.009		-0.009	*
	(0.009)		(0.005)	
Entries x Trt[None]	0.001		0.008	
	(0.010)		(0.005)	
Entries x Trt[Random]	0.007		0.009	**
	(0.007)		(0.004)	
Chi-squared test	33.4		21.4	
Mean Response	0.27		2.24	
Observations	334		243	
DF	15		9	

Probit regression on repeating entrants, with cohorts, base case is Directed feedback
Negative binomial regression on count of number of logos per entrant, base case is Directed feedback
Significance levels: * <0.10, ** <0.05, *** < 0.01
Standard errors given in parentheses

‡ Omitted control variables, significant

Cohort defined as a group of entries where maximum elapsed time between consecutive entries is 2.4 hrs
Similar results obtained without cohorts (all 544 submissions)

For model 10, Entries defined as average logos in contest across an entrant's submissions (so an agent who only submits at the beginning would be represented by a low number of entries)

6 Discussion

6.1 Summary of Findings

The hypotheses we pose in Section 3 are largely supported. Directed feedback is associated positively with the quality of entries. Quality improves with cumulative entries. Quality ratchets up with the quality of the best prior entry by others and by the agent, in the presence of directed feedback. In the aggregate, the variance in quality is not different across the three treatments. However, variance declines as the contest progresses under directed feedback. Finally, the likelihood that an agent submits multiple entries increases in the presence of directed feedback.

6.2 Implications at Contest Level

While our results are supportive of a theory of the behavior of individual agents in contests, in general an administrator is concerned primarily with the net result of a contest; how good is the *best* entry. Based on the statistical conceptual framework for our work, we can infer that increasing average quality and increasing the number of submissions from an agent will increase the quality of the best idea in a contest. These effects are associated with directed feedback. The quality of the best idea should also increase in the variance in quality of submissions. While we do not observe statistically significant differences in variance across the three treatments in the aggregate, there is a reduction in variance over time with directed feedback, which could result in lower overall performance of the contest. However, the net effect does not appear to be large enough, in this setting, to warrant omitting feedback altogether or randomizing feedback.

6.3 Limitations

We constructed our experiments to run in close temporal proximity to each other without being too adjacent within the contest sites. As a result, we began the first set of contests on a Monday, the second set on a Wednesday, and the third set on a Friday. There could be submission effects that result from the patterns of site traffic and contest timing, despite being the same length overall length. We try to control for this with both day-of-week and contest-day effects, but we could be over- or under-specifying these effects.

Given the fact that we performed this experiment with real designers instead of in a lab, we could not use a within-subjects design. The reaction of the same individual under different treatments would be interesting and potentially feasible in a laboratory study.

Finally, while the contests were constructed to be nearly identical, in order to avoid detection the challenges had to differ in details. Although we control for the product type with a fixed effect, different types of challenges could attract fundamentally different types of agents, which could bias our results.

6.4 Managerial Implications

This research is motivated by the managerial question of whether or not an administrator of an innovation contest should provide directed feedback, random feedback, or no feedback. We found that on balance, directed feedback provides higher performance than does no feedback or random feedback. Providing directed feedback to low-quality ideas might inhibit idea exploration. At a minimum, administrators should communicate to agents that they are paying attention and are engaged in the contest. This is likely to encourage repeat entries.

6.5 Research Directions

There are many additional avenues for research related to this study. We expect, for instance, that a blind contest with no visibility in terms of others' submissions would induce quite different behaviors.

A further interesting analysis would be to evaluate the designs with respect to overall uniqueness, or to the design approach, possibly using the similarity methodology of Kornish and Ulrich (2011). With similarity scores for all logos, we could analyze the effects of feedback on the diversity of approaches, and not just on the variance in quality.

We considered only two options for feedback: directed feedback, which was highly correlated with the eventual quality judgment, and random feedback, which was largely uncorrelated. It is possible that there is an optimal amount of noise introduced into the feedback to stimulate exploration of alternatives while not undermining the credibility of the administrator, and continuing to provide some information about promising directions for exploration. A study of the level of noise to be introduced would be interesting.

Analysis of overall performance of contests (i.e., the quality of the winning design) is elusive because experiments are expensive and thus samples are relatively small. An observational study of a field setting with hundreds of contests might allow us to make inferences about the effect of contest parameters on overall contest performance.

Appendix A: Sample Design Briefs – Sports Gear Set

99Designs – Crazy Comet Soccer Gear

BRIEF OVERVIEW:
We are soliciting logos for a new consumer products brand that focuses on the U.S. soccer market. Crazy Comet Soccer Gear will offer a range of soccer products (cleats, practice equipment, bags, etc.) to the continually-expanding market that plays soccer in the U.S. The logo should appeal to avid players as well as attract new players to the product line.

BRAND NAME:
Crazy Comet Soccer Gear

TARGET AUDIENCE:
We are focusing on the adult soccer market. Our target audience is college-educated U.S. consumers 18-35 years old.

REQUIREMENTS:
Develop a logo that will be most appealing to our target audience and will work in a variety of applications (product logo, website, etc.). Any color scheme is okay. The use of graphics or text (or a combination of the two) is acceptable. You will help direct our branding efforts -- we are very open to all designs and don't have any limitations on what the design can or can't include.

Final selection will be based on ratings from a panel of judges drawn from the company's target market, 18-35 year old students and professionals. The winner will be notified within one week of the contest's conclusion. The final files should have the original vector files.

CrowdSpring – Supernova Swim Wear

LET ME TELL YOU ABOUT US:
We are soliciting logos for a new consumer products brand that focuses on the U.S. swimwear market. Supernova Swim Wear will offer a range of swim suits and beachwear products. Our focus is on adult swimwear for settings that include water activities and place a premium on comfort and style.

HERE IS WHAT WE NEED:
A logo that will be most appealing to our target audience and works in a variety of settings, including website, product tags, etc. Any color scheme is acceptable, and the logo can include graphics or text or a combination of the two.

Final selection will be based on ratings from a panel of judges drawn from the company's target market, 18-35 year old students and professionals, and the winner will be notified within one week of the contest's conclusion.

OUR TARGET AUDIENCE IS:
We are focusing on the adult market. Our target audience is college-educated U.S. consumers 18-35 years old.

WE LIKE THESE DESIGNS:
We like a wide range of designs and don't have any limitations on what the design can or can't include - you get to help direct our branding efforts.

FORMATS: EPS, PSD, AI (vector based), and JPG

Appendix B: Feedback and Rating Validation

Random Feedback Rating Distribution

Rating	Prescribed Frequency	Actual Frequencies		Overall Frequency
5 stars	12.5%	26%	4%	9%
4 stars	25%	16%	28%	25%
3 stars	25%	35%	32%	33%
2 stars	25%	13%	25%	22%
1 star	12.5%	10%	11%	10%
Sample size		31	143	174

Note: Actual frequencies listed for South of the Border Salsa and Bright Bay Toys contests, respectively.

Correlation of Feedback Judge Ratings to Final Judge Ratings

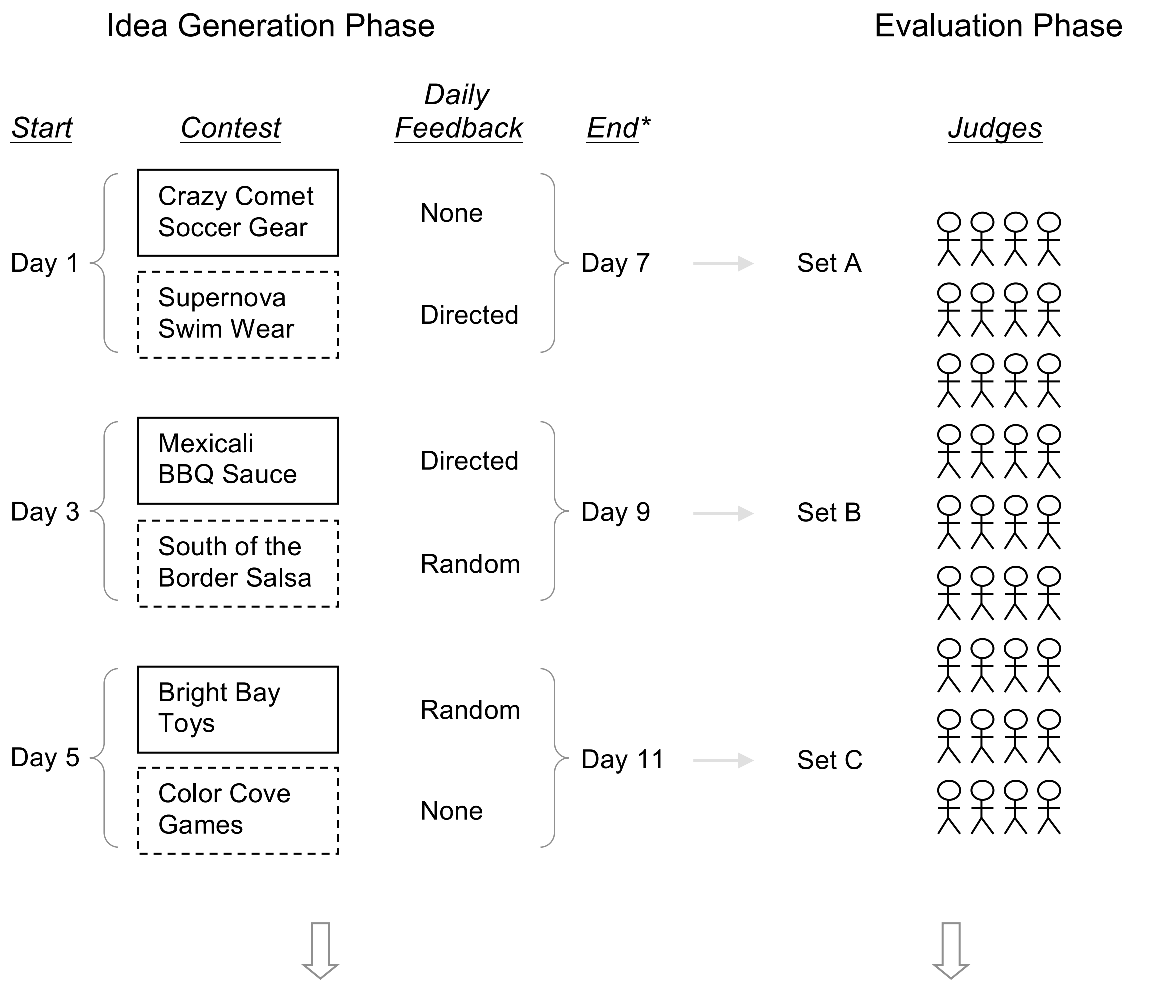
Contest	Treatment	All logos	Top-5 & Bottom-5
Crazy Comet Soccer Gear	none	–	–
Supernova Swim Wear	directed	0.73	0.98
Mexicali BBQ Sauce	directed	0.43	0.63
South of the Border Salsa	random	-0.19	-0.28
Bright Bay Toys	random	0.03	0.07
Color Cove Games	none	–	–

Inter-rater Reliability of Judges (Krippendorff Alpha)

Contest	*Ordinal Alpha* Judges' Ratings	*Interval Alpha* Logo Averages
Crazy Comet Soccer Gear	0.13	0.81
Supernova Swim Wear	0.17	0.72
Mexicali BBQ Sauce	0.16	0.75
South of the Border Salsa	0.11	0.69
Bright Bay Toys	--	--
Color Cove Games	0.16	0.73
Average	0.15	0.74

Note: Judges' Ratings analyzes degree of agreement among the 36 judges on every logo's rating; Logo Averages measures the agreement between the 36-judge panel and a second 24-judge panel on the average logo rating; Bright Bay Toys omitted because of Latin square missing values.

Appendix C: Overview of Experimental Design



99Designs (solid box)

CrowdSpring (dashed box)

None: No star rating given

Directed: 6 feedback panelists from target market segment, college-educated U.S. adults, 18-35; 1-5 scale

Random: Random number generation; 1-5 scale

36 judges from target market segment

Repeated Latin square design used to randomize the survey order for sets

Within each set, half of the judges rated the 99Designs logos first and half rated the CrowdSpring logos first

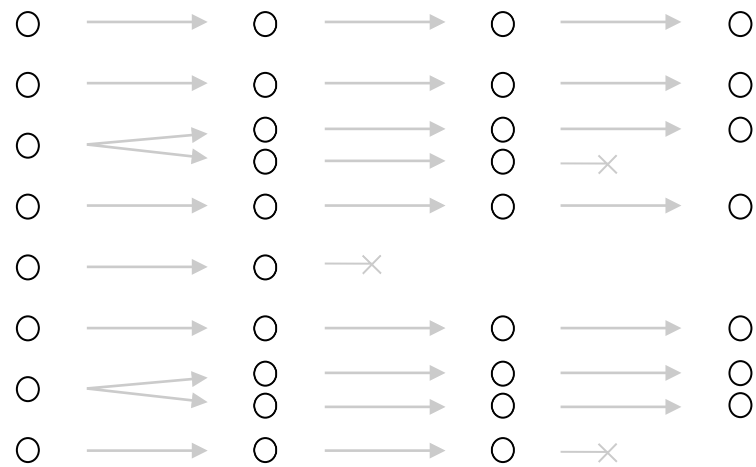
Within each survey, the logo order was randomized to the judges

544 logos rated by judges, in total

*Days are given in 24-hr blocks; each contest lasted exactly seven days (168 hours).

Appendix D: Evolutionary Path of Scored Ideas

Submissions	Total Ideas	Available	Scored
Contest entries submitted by creatives	Accounting for multiple logos per entry	Accounting for withdrawn submissions	Non-repeated entries sent to rating panel



Total #	624	635	605	544

Note: The filter from ‘Available’ to ‘Scored’ was a subjective hurdle that filtered out duplicates or near duplicates (e.g., color variations) to accomplish two goals: 1) make the judging panel inclusions non-repeated and 2) to eliminate inconsistencies in the random feedback contests.

The Impact of Visibility in Innovation Tournaments: Evidence from Field Experiments

Contests have a long history of driving innovation, and web-based information technology has opened up new possibilities for managing tournaments. One such possibility is the visibility of entries – some web-based platforms now allow participants to observe others' submissions while the contest is live. Seeing other entries could broaden or limit idea exploration, redirect or anchor searches, or inspire or stifle creativity. Using a unique data set from a series of field experiments, we examine whether entry visibility helps or hurts innovation contest outcomes. Our eight contests resulted in 665 contest entries for which we have 11,380 quality ratings. Based on analysis of this data set, we provide evidence that entry visibility influences the outcome of tournaments via two pathways: (1) changing the likelihood of entry from an agent and (2) shifting the quality characteristics of entries. For the first, we show that entry visibility generates more entries by increasing the number of participants. For the second, we find the effect of entry visibility depends on the setting. Seeing other entries results in more similar submissions early in a contest. For single-entry participants, entry quality "ratchets up" with the best entry submitted by other contestants previously if that entry is visible, while moving in the opposite direction if it's not. However, for participants who submit more than once, those with better prior submissions improve more when they can not see the work of others. The variance in quality of entries also increases when entries are not visible, usually a desirable property of tournament submissions.

1 Introduction

The key to a successful innovation tournament lies in the ability to extract the best *few* opportunities from a process that considers *many* (Terwiesch and Ulrich 2009). In such contests, participation by many agents can reduce individual effort and investment thanks to negative economic incentives (Taylor 1995, Fullerton and McAfee 1999, Che and Gale 2003) but these costs are offset by gains from the parallel search efforts of the increased number of contestants (Terwiesch and Xu 2008, Boudreau et al. 2011). This important characteristic has made tournaments effective processes for generating high quality solutions to innovation challenges (Terwiesch and Xu 2008, Terwiesch and Ulrich 2009). However, when faced with designing such contests, administrators face numerous decisions with respect to how the contest will run – from defining the challenge to soliciting entries to moderating the contest (Wooten and Ulrich 2012). Knowing that participants adapt to different incentives and information, a key managerial challenge is how a contest administrator can best design and operate a tournament.

In this paper, we examine the effectiveness of two methods of moderating entries to a contest – *blind* and *unblind*. In blind contests, an entry's visibility is limited to the individual who submitted it and the contest administrator. Without observing the work of others, agents must innovate on their own *from scratch*. In unblind contests, entries are fully visible to other participants; anyone can see the full slate of submissions. The ability to observe directly some positions in the space of possibilities means that agents no longer operate in a vacuum. Seeing other entries could broaden or limit idea exploration, redirect or anchor searches, or inspire or stifle creativity. What effect does entry visibility have on contest performance?

To answer this question, we report on a set of field experiments using web-based platforms for graphic design tournaments. We manipulate contest visibility – either blind or unblind – and use real contests and designers to test how changing the information available in the search process impacts exploration. Specifically, our goal is to test for differences in participant behavior and contest outcomes that stem from the administrator's decision about entry visibility.

Our experiment is unique in that it is the first to look at differences between innovation tournaments with varying degrees of entry visibility. The eight contests we launched resulted in 665 submissions from 224 agents over the course of a week. A panel of target consumers then rated the quality of each entry, giving us 11,380 distinct entry-ratings. Additionally, students grouped the entries into related clusters in order to quantify the similarity between submissions. These measures – along with the detailed contest administration data – allow us to analyze both participant entry and the characteristics of the work these participants submit.

Our results show that there are, in fact, differences in agent behavior and contest outcomes based on the degree of entry visibility. We find that unblind tournaments generate more entries – not by inducing more entries from existing agents but by increasing the number of agents that participate. We also find that the degree of similarity among submissions increases in early periods, provided that agents can see other entries. For single-entry participants, entry quality "ratchets up" with the best entry submitted by other contestants previously if that entry is visible, while moving in the opposite direction if it's not. Unblind contests offer an environment in which to learn about the landscape and produce better entries. However, for participants who submit more than once, those with better prior submissions improve more when they cannot see the work of others. The variance in quality of entries also increases when entries are not visible, usually a desirable property of tournament submissions.

2 Visibility in Innovation Tournaments

Innovation tournaments have been shown to be effective processes for generating novel solutions (Terwiesch and Xu 2008, Terwiesch and Ulrich 2009). In fact, they have a long history of driving progress, especially in the fields of engineering and design. Consider the famous

Tower Bridge in London, the largest and most sophisticated bascule and suspension bridge ever constructed when it went up. At that time, London's East End faced massive congestion, and delays for pedestrians and vehicles were routinely several hours. A "Special Bridge or Subway Committee" convened in 1876 and announced a contest to design a new public crossing on the Thames that wouldn't disrupt commercial river traffic; over 50 designs were submitted for consideration and produced the final design we see today.[1]

Such tournaments have typically been organized as blind contests with batched evaluation. That is, designers submit one or more entries – without knowing what other ideas are submitted – and wait for a panel to declare a winner. More recently, developments in information technology in several domains have made submission and evaluation of entries to tournaments much less costly, allowing for sequential in-process evaluation. For instance, the 2006 Netflix Prize sought a new recommendation algorithm for its movie business. By automating the judging, Netflix could provide instantaneous scoring of submissions, allowing the 5,169 teams (who submitted over 44,000 algorithms in total) to learn the quality of their entries throughout the contest and resubmit.[2] Netflix employed a blind contest with sequential evaluation – entrants still couldn't see the ideas that were submitted, but were scored in real-time and shown the distribution of results. Sequential scoring effectively changes innovation tournaments from one-shot events to environments in which individuals can participate and learn iteratively.

Information technology has enabled other modifications to traditional contest features. One such element is the blind constraint. Rather than maintain the precedent of restricting entry visibility, some platforms have pulled back the curtain, allowing entrants to see the work of other contestants. This raises the question of whether seeing other submissions helps or hurts contest outcomes. Anecdotally, the market believes visibility of entries influences outcomes. The web-

[1] The Corporation of London and the Tower Bridge Exhibition, www.towerbridge.org.uk (2013)
[2] Netflix Prize, www.netflixprize.com (2013)

based contest platform 99Designs, one of the sites we use in our field experiments, advertises that blind contests attract better designers, promote creativity, and result in higher quality entries.[3] However, there has been no prior empirical evidence that has directly explored this impact of entry visibility in innovation tournaments.

3 Literature and Hypothesis Development

Before we develop our hypotheses, we categorize some of the current literature in terms of the type of contest examined (Table 1). This classification is used throughout the rest of the paper and, more generally, as a review of the research that directly deals with innovation contests.

In this paper, we examine the impact of moderating entry visibility by looking at blind and unblind contests. In blind contests, an entry's visibility is limited to the individual who submitted it and to the contest administrator; other participants may see ancillary information – such as who submitted it or the rating it received – but not the innovation itself. This requires agents to innovate on their own. In unblind contests, submitted entries are fully visible to other participants; anyone can see the full slate of submissions.

In what ways might visibility of entries alter tournament outcomes? Once an agent has committed to join a tournament, visibility of other entries should theoretically be beneficial to his or her problem solving efforts. The other entries can be viewed simply as additional information – and from that perspective should not degrade an agent's performance relative to not having that information. Indeed, an agent could simply ignore the other entries and work from the problem statement with no other information. The agent could then consider the other entries, and decide whether or not to create additional entries based on that newly available information.

While theoretically appealing, this argument may not reflect the realities of human behavior. People are unlikely to actually ignore readily visible entries from rivals, especially as they

---

[3] 99Designs, www.99designs.com (2013)

Table 1. Tournament Characteristics within the Literature

Paper	Methodology	Visibility	Evaluation	Key Result
Bayus (2012)	Observational study (IdeaStorm)	Unblind	Sequential	Ideators can get stuck, generating less diverse ideas after success
Poetz & Schrier (2012)	Natural experiment (company)	Blind	Batched	Users in the market can outperform professionals in some ideation settings
Wooten & Ulrich (2012)	Field experiment (99Designs, CrowdSpring)	Unblind	Sequential	Seeing feedback increases entrants and quality; Quality rachets up with prior best
Bockstedt et al. (2011)	Observational study (LogoMyWay)	Unblind	Sequential	Winning agents enter earlier, submit over time (but aren't helped by add'l entries)
Boudreau et al. (2011)	Observational study (TopCoder)	Blind	Sequential	Uncertainty amplifies the benefit of parallel search
Erat & Krishnan (2011)	Analytical	Blind	Batched	Searchers cluster in regions (search breadth increases with entrants sub-linearly)
Kornish & Ulrich (2011)	Lab experiment (university)	Blind	Batched	Incidence of redundancy in parallel search is small (<13% across several domains)
Walter & Back (2011)	Observational study (Atizo)	Unblind	Batched	Increases in rewards and market maturity attract more submissions
Bullinger et al. (2010)	Observational study (university)	Unblind	Sequential	Cooperation amongst agents can be a beneficial search strategy
Jeppesen & Lakhani (2010)	Observational study (InnoCentive)	Blind	Batched	Agents from outside fields have a higher chance of winning
Y. Yang et al. (2010)	Observational study (TaskCN)	Blind	Sequential	Higher award levels, less crowded markets, and easier tasks lead to more entries
Archak & Sundararajan (2009)	Analytical	Blind	Batched	All significant outcomes will be determined by the most efficient contestant types
J. Yang et al. (2008)	Observational study (TaskCN)	Blind	Sequential	Average users do not increase their chance of winning with experience
Terwiesch & Xu (2008)	Analytical	Blind	Batched	More entrants and parallel search offset underinvestment by any one contest agent
Lakhani et al. (2007)	Observational study (InnoCentive)	Blind	Batched	Winning agents work twice as long on their solutions
Soukhoroukova et al. (2007)	Field experiment (company)	Unblind	Sequential	Group decisions in visible idea markets can improve evaluation
Toubia & Flores (2007)	Sim, Field experiment (company)	Blind	Batched	Ideas can be misclassified or misjudged by users
Dahl & Moreau (2002)	Lab experiment (university)	Blind	Batched	Analogical thinking can benefit ideation strategies
Gradstein & Konrad (1999)	Analytical	Blind	Batched	Multi-stage contests differ from one-shot events

consider whether or not to join a tournament. Thus, the visibility of the entries of others is likely to influence the outcome of a tournament in at least two basic ways. First, the visibility of the entries of others may influence the likelihood of entry from an agent, altering the number of entrants, their composition, and number of entries each submits. Second, the visibility of the entries of others may influence the way in which a particular agent addresses the challenge, possibly leading to differences in the search process and quality of entries submitted by that agent. We refer to these two pathways of influence as *entry* and *characteristics of entries*, respectively. We consider each of these in turn, relating the effects to the literature and posing hypotheses for our experiments.

3.1 Entry

The number of entries to a tournament is a function of both the number of entrants to the tournament and the number of entries submitted by each entrant. Here we consider how entry visibility impacts each of these variables.

The number of contest entrants could increase with entry visibility because of a lower cost of entry, more appealing community experience, and from a superiority bias on the part of entrants, or the number of entrants could decrease as a result of intellectual property concerns.

Individuals might face lower entry costs thanks to having a better map of the solution landscape, lots of seed ideas from which to begin their search, or exemplars that can be changed incrementally with less work than starting from scratch. In searching for solutions, effective strategies can include analogical thinking (Dahl and Moreau 2002), recombination of acquired expertise (Lakhani et al. 2007), and cooperation among agents competing in the same search (Bullinger et al. 2010). This idea is partly formalized as the path of least resistance, an idea within psychology's *structured imagination* construct where people modify existing solutions when faced with problems requiring creativity (Ward 1994). We see one derivation of this idea from

Wooten and Ulrich (2012), in which knowing where good ideas occur on the landscape – through visible feedback – results in more contest entrants over time.

Increased visibility could promote more appealing social engagement and intrinsically more interesting work. Seeing other entries could also trigger cognitive biases (Alick et al. 1995) and induce greater participation from a better-than-average self-perception. Each of these suggests that seeing other entries may results in more entrants per contest.

Bockstedt et al. (2011) highlight one disadvantage of entry visibility; namely, the perceived potential for intellectual property loss. If the perceived threat of having an idea “stolen” is high enough, it could be a deterrent to entry. Of course, instead of opting out, agents could decide to devote more effort and stake a claim to the area around an idea, with increased submissions to discourage infringement from competitors, which leads to our second participation variable – entries per entrant.

The number of entries per entrant could increase with entry visibility thanks to lower search costs – in much the same way as the entry decision could be affected. It is possible that a spirit of competition is induced by revealing the work of the participants. In unblind contests, several empirical studies analyze how contest characteristics impact contestant participation (Table 1), including increased entries with market maturity (Walter and Back 2011) and less complex tasks (Yang et al. 2010). However, most of these studies study total contest entries instead of the behavior of contestants within a contest. Bockstedt et al. (2011) empirically demonstrate that winning agents on LogoMyWay.com are more likely to enter earlier and submit entries over a wider range of time, but aren’t helped by simply entering more ideas.

On balance, we expect that greater entry visibility in innovation tournaments will result in increased participation. All but one of the hypothesized effects suggest that contest entries will be greater in contests with entry visibility. Unblind contests make an agent’s key decisions easier. The choices around whether to enter and the amount of effort to invest both derive benefits from entry visibility. By seeing other entries in the landscape of possibilities, the barriers to entry are lower for any given agent and more information on the administrator's quality function is available. Easier search should result in more entry.

Hypothesis 1: Increasing entry visibility in an innovation tournament (by moving from a blind to an unblind contest) will increase the number of entries submitted.

3.2 Characteristics of Entries

Given that we expect the number of entries to change, do the characteristics of those entries also change? Entry visibility may influence the way in which a particular agent addresses the challenge, possibly leading to differences in the search process and quality of entries submitted by that agent. Two relevant metrics of the characteristics of entries are *similarity* and *quality*, including both the mean and distribution.

Similarity. Independent of idea quality, seekers usually benefit from knowing the landscape – observing diverse ideas gives a more complete picture of the solution possibilities. The incidence of redundancy in parallel search has been shown to be quite small in blind contests (Kornish and Ulrich 2011). In unblind contests, entry visibility could mean even less redundancy in effort, with agents using the knowledge of competitors' submissions to reduce repetition. Or such visibility could inhibit parallel search, with entrants clustering their submissions around existing proven entries (Erat and Krishnan 2011). Either way, if a participant searches differently in response to seeing other entries, then the resulting similarity among entries should change.

In a set of graphic design prototyping experiments around online ads, participants who saw multiple shared designs borrowed significantly more features to incorporate in their own ads (Dow et al. 2012). In creativity tasks, Marsh et al. (1996) found that individuals who saw many examples tended to incorporate critical elements in their own designs (although without inhibiting creativity), and Smith et al. (1993) found conformity in every group that saw examples, across a range of conditions and instructions. In unblind contests, more designs will be visible to agents, and we expect the prior conformity results to play out in innovation contests.

Hypothesis 2: Increasing the visibility of entries in an innovation tournament will result in submissions that are more similar.

Quality. At the level of the contest, the population of entries yields a distribution of quality, reflecting the overall performance of the tournament. This idea arises from the statistical view of innovation processes (March 1991; Dahan and Mendelson 2001; Terwiesch and Ulrich 2009). One way to describe the quality distribution is with mean and variance, and increases in each of these variables improve the overall performance of tournaments (Girotra et al. 2010).

The mean quality of entries is driven by both the quality of entrants and the quality of the work those entrants do. If a tournament attracts better entrants or better submissions from its existing entrants, overall contest performance improves. However, in many settings, it's not possible to truly disentangle the intrinsic quality of entrants from the work they do. Here, we rely on entry quality as the aggregate measure of these two drivers and explore how that quality might be influenced by entry visibility.

Exposure to additional information in unblind contests likely impacts the learning environment. Openness and information sharing has long been identified as important to scientific progress (Merton 1942, Mulkay 1975), with examples such as open source software development at the recent forefront (von Hippel 2005). In evolutionary economics, the role of search has been highlighted as a mechanism for discovering variety and allowing organizations to develop new technologies (Nelson and Winter 1982). Metcalfe (1994) suggests that exploring such variety allows firms to innovate more successfully by seeing a range of potential options or paths to explore.

We would expect participants to learn the most and have the best understanding of the search landscape when full information from all the parallel searches is visible. In the design world, having examples readily available has been shown to improve the appeal of designs, although these benefits appear to accrue to novice designers more than to experts (Lee et al. 2010).

Some operators of web-based platforms for innovation contests assert that blind contests result in better entries, with the rationale that blind contests attract higher quality talent. If better designers don't benefit from the presence of examples as Lee at al. find, then other benefits of the blinded format (such as intellectual property protection) could be attractive. On balance, we believe that there is more evidence on the side of increased information and learning, as mechanisms for increasing the average quality of entries.

> *Hypothesis 3*: Increasing the visibility of entries in an innovation tournament will increase the average quality of entries submitted.

Finally, variance in the quality of submissions, for a given mean, improves tournament outcomes, as flatter distributions result in more ideas in the upper tail of the distribution (Girotra et al. 2010). Such benefits could be driven by both variance in the quality of entrants and by variance in the quality of the work they do. Given the uncertainty in the task and conditional on a given set of entrants, variance in approach is expected to be one of key drivers of variance in quality. The way in which an agent searches the landscape likely impacts variance in the quality distribution. Thus, it follows from our similarity hypothesis (H2) that we expect less variance in approach in visible tournaments, and by implication less variation in quality. Wooten and Ulrich (2012) similarly found that more information about the administrator's quality function results in a convergence of approaches and decreased variance in the quality of contest submissions.

> *Hypothesis 4*: Increasing the visibility of entries in an innovation tournament will decrease the variance in quality of entries.

4 Experimental Design

We conducted a set of field experiments in which we explicitly control the environment and compare the performance of innovation contests with varying levels of visibility. We've used these platforms for experiments before, however, here we use a completely new set of experiments designed specifically to address the issue of visibility in contests. We follow similar conventions as those used by Wooten and Ulrich (2012) for the setup, delivery of feedback, and measure of entry quality in an online graphic design field experiment. Four pairs of logo design competitions were posted on two online design contest marketplaces, 99Designs and CrowdSpring. The competitions differed in terms of the amount of information visible to entrants – in the unblind treatment, agents could see all entries and feedback while in the blind treatments, the entries of others were not visible. At the conclusion of the contests, a consumer panel rated the quality of each entry and a pool of university students rated their similarity.

4.1 Contest Platforms

Our experiments were hosted by two online companies, 99Designs and CrowdSpring, that have emerged as leaders in the crowd-sourced design market. They allow buyers to solicit projects – such as logo creation – from a community of graphic designers. While buyers are mostly small businesses and entrepreneurs, established companies such as Amazon, Starbucks, Microsoft, Philips, Barilla, and TiVo have also run contests. Contest winners are awarded predetermined cash prizes – normally between \$150 and \$1,500 per contest. The sites support robust marketplaces. As an example, 99Designs has awarded over \$43 million worth of contest prizes in more than 174,000 contests since its founding in 2007.

The two platforms are very similar, with nearly identical interfaces and business implementations. Each website counts over 125,000 designers as members and targets an array of design projects (such as logos, packaging, book covers, and website design). Clients create a contest by posting project specifications and a prize amount. Over a project's duration (generally one week), online submissions are submitted by interested designers and feedback can be given by the client.

4.2 Contests

Four pairs of contests were launched as follows.

A: Burning Barn BBQ Sauce Smoking Silo Salsa	B: Wave Monkey Headphones Sound Chimp Speakers
C: Power Perk Coffee Bold Brew Tea	D: Jailbird Dog Gear Rat Pack Cat Company

All eight contests had similar details, and within each pair, projects had nearly identical details, including company type, name, design specifications, deliverables, target markets, and specifics of the design brief. Each logo was for a new consumer product brand whose target audience was specified to be college-educated U.S. consumers 18-35 years old. Designers were told that a panel of consumers from this market would be the ultimate judges of entry quality. The contests in each pair shared the type of product (condiments, audio electronics, beverages, and pet accessories), were constructed with similar name characteristics and motifs, and were randomly assigned to one of the two websites. An example of the submitted design briefs is Appendix A.

Designers count on feedback over the course of contests to determine performance of any particular entry. The established feedback mechanism on both 99Designs and CrowdSpring is a one-to-five star rating, which indicates how much the administrator likes an entry. We provided new entries with feedback every morning using this scale; a three-person panel of independent judges scored each design and their average determined the rating, expressed as a number of stars. The raters fit the target market demographic (consistent with our design brief), and we used two such panels to manage the volume from four concurrent contests. This feedback was intended to be highly correlated with the final ratings which would eventually be produced by an evaluation by a larger panel of consumers at the conclusion of the contest.

4.3 Treatments

The independent variable tested was entry visibility in each contest. CrowdSpring and 99Designs permit both blind and unblind contests, which allows the administrator to choose at the outset who can see a designer's submissions. In unblind projects, anyone who views the contest can see the full slate of designs that have been entered as well as any scored feedback given (in the form of star ratings). Thus, the general public has full information about submissions and their in-process ratings. In blind projects, an entry's visibility is limited to the designer who submitted it and the contest administrator. Other designers know how many designs have been entered – and by whom – but are restricted from viewing the actual submission.

4.4 Experiment

We denote the four pairs corresponding to the four product types as A, B, C, and D. One of each pair ran on 99Designs and its nearly identical corollary ran on CrowdSpring, allowing for each visibility treatment to be tested twice on each site in a balanced design. Designers closely monitor the contests on these websites and frequently report copyright violations and other such concerns. To deal with such savvy agents and avoid undermining the outcomes, we constructed the experiment design to utilize two different website platforms, slightly staggered start dates, and small differences in the award levels. The contest pairs ran over the course of two weeks. Sets A and B ran during the first week, and sets C and D ran during the second. CrowdSpring and 99Designs display the award amount in slightly different ways, but sets A and C carried award levels of \$250 for the winner, and awards for sets B and D were \$237. These slight differences were built into the contest setup to make the contests nearly identical, without tipping the designers off that the products weren't real. The visibility treatment can be denoted by subscripts (*B* for blind and *U* for unblind) resulting in the following contest layout:

99Designs:	A_B	B_U	C_U	D_B
CrowdSpring:	$\bar{A}_U$	$\bar{B}_B$	$\bar{C}_B$	$\bar{D}_U$

The eight contests relied on the standard mechanisms of the websites to entice designers to participate. Each contest was open to anyone on the respective website, ran for seven days, received daily feedback, and resulted in an award to the winning designer. All experiments were conducted after obtaining approval from the human subjects committee at the university.

4.5 Evaluation

A total of 665 entries were generated by 224 designers over the course of the eight tournaments. Two panels of 20 judges independently and anonymously evaluated the logos from the perspective of potential consumers. The judges were representative of the target market outlined in the contest briefs – college-educated individuals between the ages of 18 and 35. These judges were similar in profile but distinct from the feedback panelists, who provided the daily star ratings.

Ratings were collected using web-based surveys. One panel of judges rated logos in sets A and D; the other rated logos in sets B and C. Following the design of Wooten and Ulrich (2012), entries from the eight contests were administered in separate surveys and were completed as paired sets. To mitigate order effects, surveys were administered as a balanced, repeated Latin square design; each set order (AD, DA, BC, CB) appeared the same number of times. Within each set, half the judges were given the contest from 99Designs first followed by the one from CrowdSpring; the other half saw them in the opposite order. Within each individual survey, the logos were presented to each judge in a randomized order. The question and response choices (on a 1-5 rating scale) were the same for the judges as for the in-contest feedback panel.

One contest (Wave Monkey Headphones, with 192 entries) exceeded the survey length threshold established in similar settings (Girotra et al. 2010, Wooten and Ulrich 2012). As a result, each judge rated half of the designs in that contest; for those 20 judges, the assignment of particular logos followed a balanced, repeated Latin square design in which each rater saw 96 logos and each logo received 10 raters.

The judges' responses to the eight surveys provide the measure of entry quality for our analysis. We find that the reliability of judges is high. We check this using a Krippendorff alpha test on our population of raters (Table 2). Given the artistic nature of our contests, we expect high variation in the scores because of personal preferences. This is corroborated with a relatively low degree of agreement between any two judges. However, if populations have stable preferences, then a high degree of agreement should be seen in the average scores of entries across populations. We test this with a bootstrap approach, splitting our judges into two randomized groups and comparing the average scores for each logo between groups. With this population-level approach, we obtain an agreement alpha of 0.72, above accepted thresholds. A sample of the scored logos is provided in Figure 1.

Table 2. Inter-rater Reliability of Judges (Krippendorff Alpha)

Contest	*Interval Alpha* Judges' Ratings	*Interval Alpha* Bootstrap Averages
A: Burning Barn BBQ Sauce	0.35	0.84
B: Smoking Silo Salsa	0.21	0.80
D: Sound Chimp Speakers	0.04	0.72
E: Power Perk Coffee	0.12	0.72
F: Bold Brew Tea	0.11	0.65
G: Jailbird Dog Gear	0.23	0.74
H: Rat Pack Cat Company	0.08	0.57
Average	0.16	0.72

Note: Judges' Ratings analyzes degree of agreement among the 20 judges on every logo's rating; Bootstrap Averages measures the agreement between a 10-judge random sample of the 20-judge panel and the remaining 10 judges on the average logo rating; Wave Monkey Headphones omitted because of Latin square missing values.

Figure 1. Examples of Logos Generated (sets A and B)

	Highest Rated	Median	Lowest Rated
Burning Barn BBQ Sauce [1,B]			
Smoking Silo Salsa [2,U]			
Wave Monkey Headphones [1,U]			
Sound Chimp Speakers [2,B]			

1: 99Designs, 2: CrowdSpring; B: Blind (no visibility), U: Unblind (full visibility)

5 Data

Table 3 is a summary of each contest's outcome. Given our experimental design, we also have a great deal of entry-level data, which we analyze to test our four hypotheses.

5.1 Measuring Entry

To measure entry behavior we capture *entries*, *entrants*, and *entries per entrant*.

An *entry* is defined as an idea submission to a particular contest and captures the aggregate level of participation in a contest. The more participation a contest elicits, the more entries there are, resulting in more potential solutions for the contest administrator.

An *entrant* is a distinct contest participant, someone who submits at least one entry. The more attractive the contest, the more entrants it attracts, which increases the number of parallel searches that occur.

Entries per entrant is defined as the number of submissions by a contest participant. We use

Table 3. Contest Summary

	Condiments		**Audio Electronics**		**Beverages**		**Pet Accessories**	
	Burning Barn BBQ Sauce	Smoking Silo Salsa	Wave Monkey Headphones	Sound Chimp Speakers	Power Perk Coffee	Bold Brew Tea	Jailbird Dog Gear	Rat Pack Cat Company
Visibility	Blind	Unblind	Unblind	Blind	Unblind	Blind	Blind	Unblind
Website	99D	CS	99D	CS	99D	CS	99D	CS
N Entries	82	57	192	53	91	69	40	81
N Entrants	36	24	48	20	30	29	8	29
Best Logo	3.75	3.65	3.70	3.26	3.15	3.20	3.35	3.40
Mean Logo	2.54	2.44	2.40	2.32	2.21	2.51	2.36	2.45
S.D.	0.77	0.58	0.50	0.35	0.48	0.45	0.63	0.39
Average Similarity	0.186	0.196	0.217	0.161	0.186	0.164	0.170	0.210

Note: Values listed for logos are averages of judges' ratings on 1-5 scale; Wave Monkey scores adjusted for judge.

it to estimate the effort invested by an entrant with the idea that submitting more entries requires additional effort.

5.2 Measuring Characteristics of Entries – Similarity

The second question we ask concerns the search process: How does the ability to see other entries change the way in which agents address the challenge?

To assess whether agents incorporate elements from previously submitted entries, we need a quantitative measure of logo similarity. Kornish and Ulrich (2011) tackle a similar problem in rating sets of innovation opportunities. We adopt a similar methodology in order to obtain a similarity score for every pair of entries in a contest. We had student subjects in the university behavioral laboratory form groups of similar entries from a packet of logo submissions. A packet contained a subset of logos from a single contest; each logo was printed on a square of cardstock. This allowed students to visually sort and re-sort the logos into piles quickly. Entries could be categorized into more than one cluster.

We created 45 such packets, with overlapping subsets of entries such that most entry pairs appeared multiple times. When multiple logos varied only by color, we only included only one version of a logo. In total, we ran 89 students through our protocol. They were paid $5 for participating. Sessions were not timed and most students finished the grouping task in 10-20 minutes. The grouping task resulted in a list of idea clusters that we could turn into a measure of pairwise similarity. The average cluster contained 5.5 logos per group. We coded each of the entries grouped together as *similar* and calculated an overall score between every possible pair based on the percentage of times those two entries were placed in the same cluster. This measure is the number of times two logos were grouped together over the number of times such a pairing was possible. The final score is modified to account for our packet structure and the subsets included. The similarity score between any two entries i and j is represented in the matrix A_{ijk}, where k represents the contest.

To measure how changes in entry visibility affect the similarity of submissions, our similarity metric takes two forms. *Average contest similarity* is the mean of all possible pairwise similarities within a contest, $\bar{A}_k$. *Logo-level similarity* is the similarity score of a particular logo based on all the logos submitted before it. In other words, how similar a particular logo j is to prior logos i, $\bar{A}_{ijk}$ where $i < j$, for each k.

5.3 Measuring Characteristics of Entries – Quality

To measure contest quality, we operationalize the quality distribution through two parameters – mean and variance. One benefit of this approach is that it breaks the measure of quality into underlying variables and helps mitigate the problem of sampling only winning ideas, which can be noisy in a small sample of contests.

For the *quality* models, the quality measure comes from the judges' scoring of contest entries and the unit of analysis is the rating of a particular judge of a particular logo. Individual logo rating is thus the dependent variable with the following independent variables: treatment (Blind), cumulative number of entries (Entries), the best prior submission by others (Max), and the agent's best prior entry (MaxI).

For variance in the quality of ideas, we construct a measure of *variance* for the dependent variable that takes out the linear quality improvement trend over the course of the contests, identical to Wooten and Ulrich (2012) and similar to Girotra et al. (2010). Table 4 provides descriptive statistics and correlations for the variables used in our analysis.

Table 4. Descriptive Statistics and Correlations

	Variable	Definition
(1)	Rating	Numerical score of quality for an idea from judge(s)
(2)	Visibility (blind)	Contest visibility treatment – 0: Unblind, 1: Blind
(3)	Domain	Control for product area – 0: Condiments, 1: Audio elec., 2: Beverages, 3: Pet prod.
(4)	Site	Control for platform – 0: 99Designs, 1: CrowdSpring
(5)	Day	Control for day of week – 1: Mon, 2: Tue, … 7: Sun
(6)	Judge	Control for individual providing the rating
(7)	# Prior entries	Number of entries submitted to a contest at a logo's time of entry
(8)	Max prior qlty (others)	Highest score produced by others in contest thus far
(9)	Max prior qlty (own)	Highest score produced by a given entrant's agent in contest thus far
(10)	# Entries	Number of submissions in a contest (or over a specified time)
(11)	# Entrants	Number of unique participants who submit at least one entry
(12)	# Entries/entrant	Number of entries submitted by each entrant in a contest
(13)	Avg. contest similarity	Mean of all possible pairwise similarity ratings from survey panel in a contest

Logo-level correlations (11,380 observations):

Variable	Mean	St. dev.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Rating	2.41	1.18								
(2) Visibility (blind)	0.37	0.48	0.03							
(3) Domain	1.40	1.01	-0.03	-0.15						
(4) Site	0.39	0.49	0.02	0.07	0.18					
(5) Day	5.09	1.89	-0.03	0.03	0.00	-0.00				
(6) Judge	22.68	11.34	-0.03	-0.06	0.05	-0.12	0.05			
(7) # Prior entries	52.77	44.66	-0.02	-0.31	-0.1	-0.3	0.58	0.23		
(8) Max prior qlty (others)	3.04	0.96	0.03	0.06	-0.11	0.05	0.42	-0.18	0.4	
(9) Max prior qlty (own)	1.75	1.32	0.07	-0.07	0.08	-0.08	0.11	0.01	0.12	-0.09

Contest-level correlations (8 observations):

Variable	Mean	St. dev.	(1)	(2)	(3)	(4)	(10)	(11)	(12)
(1) Rating	2.41	0.04							
(2) Visibility (blind)	0.50	0.19	0.26						
(3) Domain	1.50	0.42	-0.30	0.00					
(4) Site	0.50	0.19	0.22	0.00	0.00				
(10) # Entries	83.13	16.67	0.06	-0.50	-0.18	-0.41			
(11) # Entrants	28.13	4.16	0.31	-0.42	-0.41	-0.24	0.87		
(12) # Entries/entrant	3.06	0.34	-0.33	0.01	0.52	-0.56	0.19	-0.28	
(13) Avg. contest similarity	0.19	0.01	0.16	-0.83	-0.10	-0.18	0.70	0.65	0.07

6 Analysis and Results

6.1 Entry

Our main variable of interest is *entry visibility*, denoted in our experiments as either blind (low visibility) or unblind (high visibility). To understand how differences in entry visibility affect agent behavior, we estimate variations of the model:

$$Y_i = \alpha + \beta(\textit{Entry Visibility})_i + \delta_i + \varepsilon_i \,.$$

The dependent variable Y varies over the contests i and takes on one of the outcome variables discussed above (entries, entrants, entries per entrant). Since these measures are all counts, our model assumes a negative binomial distribution[4], which adds an over-dispersion parameter and is generally more conservative than estimates with a Poisson count model (Hilbe 2011). To control for differences across contests, which could influence our behavior measures if not accounted for, we include several fixed effect controls (δ_i) for the *domain*, *site*, and *day*. Table 4 provides variable details and gives descriptive statistics and correlations for the variables used in our analysis.

Table 5 shows the results of our negative binomial regression analysis around entry. We begin by estimating the baseline model (column 5-1) by relating *entries per contest* to *entry visibility* and including our contest fixed effects – *domain* and *site*. We find that increasing visibility (from blind to unblind) results in a significant increase in number of entries for a contest. The magnitude of this effect is over 39 additional entries per unblind tournament[5] – a substantial 60% increase from blind cases. Because our contests occur over time, we extend the model to include *contest day* as an explanatory effect and *entries per day* as the dependent variable (column 5-2). The coefficient observed for *contest day* is positive and significant,

---

[4] In assuming a negative binomial distribution for our dependent variable counts, our model includes a log link, and the resulting log-linear function can be represented as $\ln(Y_i) = \alpha + \beta(\textit{Entry Visibility})_i + \delta_i + \varepsilon_i$.

[5] Given by exp(4.647) – exp(4.647-0.472) = 39.2

showing that more entries arrive at the end of contests, which matches our experience with this domain and these platforms in the past. These results mirror our baseline model, with significantly more submissions in unblind contests.

Increased entries in unblind contests could stem from attracting more entrants or from enticing existing agents to submit more ideas, as outlined in section 3.1. Our participant models (columns 5-3 and 5-4) address the first alternative, with negative binomial regressions using *entrants per contest* and *entrants per day* as the dependent variables. In both cases, more agents are choosing to participate in unblind contests. The magnitude of this effect is about 12 more entrants per unblind contest.[6] If we look specifically at *entries per entrant* across the contests (column 5-5), we see no differences in behavior, with 225 agents submitting on average 2.96 entries per contest regardless of entry visibility. This resonates nicely with the prior finding that submitting extra entries in unblind contests doesn't increase an agent's chance of winning (Bockstedt et al. 2011).

These results support Hypothesis 1. Unblind contests generate more entries; however they do this *not* by inducing more entries from existing agents but by increasing the number of agents that participate. An interpretation of this result is that entry visibility reduces barriers to entry, allowing easier exploration of the search landscape and enticing more agents to search for a solution and submit.

[6] Given by exp(3.667) – exp(3.667-0.371) = 12.1

Table 5. Comparison of Contest Productivity between Visibility Treatments

	5-1	5-2	5-3	5-4	5-5
Dependent variable	Entries per contest	Entries per day	Entrants per contest	Entrants per day	Entries per entrant
Explanatory variables	Contest fixed effects	Contest day	Contest fixed effects	Contest day	Contest fixed effects
Constant	4.647 ***	1.472 ***	3.667 ***	1.048 ***	0.970 ***
	(0.225)	(0.217)	(0.265)	(0.221)	(0.143)
Treatment					
Blind	-0.472 **	-0.461 ***	-0.371 *	-0.387 ***	-0.078
	(0.185)	(0.135)	(0.225)	(0.132)	(0.124)
Fixed effects	Yes	Yes	Yes	Yes	Yes
Timing					
Day		0.273 ***		0.229 ***	
		(0.033)		(0.034)	
Chi-squared test	9.6	56.4	5.1	50.0	14.7
Mean Response	83.10	12.09	28.13	5.98	2.96
Observations	8	55	8	55	225
DF	5	6	5	6	5

Negative binomial regression on contest productivity counts, base case is Unblind visibility
Significance levels: * <0.10, ** <0.05, *** < 0.01
Standard errors given in parentheses

Note: 5-4 reports on every unique daily entrant; robust to excluding entrants who submit on prior days.
Mean Response listed is log-transformed to show actual values for dependent variable measure.

6.2 Characteristics of Entries – Similarity

Having shown that the entry decision varies with entry visibility, we now turn to how that behavior impacts the search process. We use our pairwise similarity measures (from section 4.2) to determine whether designers create submissions that are more similar when they are permitted to see others' entries.

First, we examine the contests at an aggregate level by comparing the *average contest similarity* with a simple t-test on the means (Table 6). Average contest similarity is the mean of all possible pairwise similarities within a contest. This captures, independent of when logos were submitted, how alike our lab group believed a contest's entries to be. We find that in aggregate, average similarity in unblind contests is approximately 14% greater and significant (0.194 vs. 0.170; t-statistic 2.59). This is meaningful, and supports Hypothesis 2, but to better capture the

Table 6. Comparison of Idea Similarity within Contests

Overall contest similarity score (mean of all contest pairwise scores):

	Blind	Unblind
Average Pairwise Similarity	0.170	0.193
Number of observations	4	4

T-statistic: 2.59 **

Significance levels: * <0.10, ** <0.05, *** < 0.01 (two-tailed test)
Overall contest similarity score is mean of each pairwise score in a contest.
Robust to more conservative measures (i.e., omitting any pairs from the same designer)

Logo-level similarity scores:

	6-1	6-2	6-3
Dependent variable	Similarity to prior entries	Similarity to prior entries	Similarity to prior entries
Explanatory variables	Contest fixed effects	Interaction with day	Interaction with period
Constant	0.230 ***	0.406 ***	0.724 ***
	(0.016)	(0.027)	(0.037)
Treatment			
Blind	-0.038 ***	-0.158 ***	-0.529 ***
	(0.014)	(0.039)	(0.058)
Fixed effects	Yes	Yes	Yes
Timing			
Day		-0.034 ***	
		(0.004)	
Day x Trt[Blind]		0.024 ***	
		(0.007)	
Period			-0.509 ***
			(0.035)
Period x Trt[Blind]			0.506 ***
			(0.058)
R-squared	0.03	0.12	0.27
Mean Response	0.21	0.21	0.21
Observations	633	633	633
DF	5	7	7

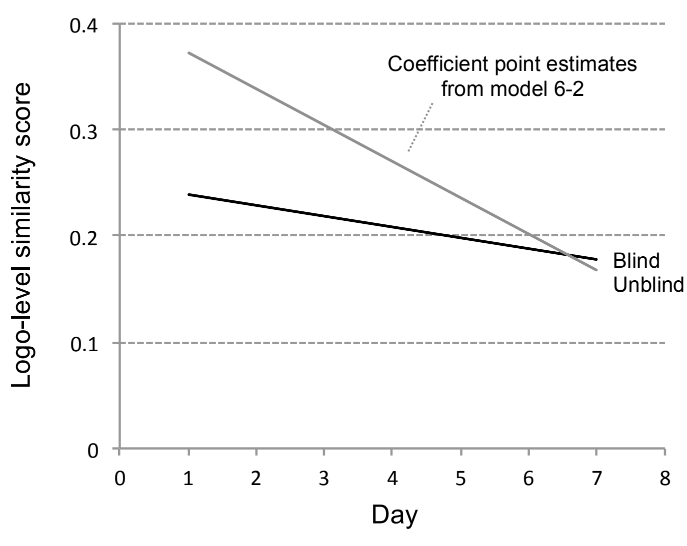
OLS regression on idea similarity scores, base case is Unblind visibility
Significance levels: * <0.10, ** <0.05, *** < 0.01
Standard errors given in parentheses

Logo-level similarity score is mean similarity of each entry to prior entries. Period is binary and defined as the first two days (0) or days three through seven (1). Robust to only first entries.

degree to which agents are incorporating elements from prior designs, we extend our analysis.

Model 6-1 shows the baseline results of our linear regression for submission similarity. Our dependent variable is *logo-level similarity*, which for each entry is the degree of similarity to *prior* submissions. We find that increasing visibility (from blind to unblind) results in entries that are significantly more similar. The magnitude of the effect is such that unblind contests were rated as 20% more similar. Including time effects (column 6-2), however, notable differences emerge. Figure 2 highlights that while entries in unblind contests are much more similar initially, by the final day, that difference has been erased. At that point, entries to unblind contests are just as unique as those in their blind counterparts. Probing a bit further (column 6-3), we can use period categorical variables to see that the difference between blind and unblind contests in terms of entry similarity happens almost exclusively in the first two days of our experiments. After that, there is no discernible difference between the treatments.

Figure 2. Entry Similarity over Time



The implication is that while unblind contests do encourage submissions that are more similar, that phenomenon is limited to the early stages of the contest. Several things could be happening. This could be because participants only incorporate elements from prior entries early in the process. More likely, once there is a sufficient breadth of entries, inspiration will have more seeds from which to spring and the resulting conformity will be harder to detect. This could be the result of a diffusion process, in which an initial seed is planted and ideas radiate out from that seed. As the ideas radiate out into a larger area, there are a greater number of seeds from which to create an incremental variant and average similarity declines. This explanation is plausible, given the results over the course of the contest. The data suggests that by increasing the visibility within tournaments, resulting submission are more similar, but that this effect quickly disappears. On balance, it appears to not overwhelm the pool of entries with conformity, which is beneficial from the administrator's standpoint.

6.3 Characteristics of Entries – Quality

Table 7 shows the results of a regression analysis with logo *rating* as the dependent variable; our explanatory variables and contest fixed effects (section 4.3) are also included. We use a clustered OLS because there are multiple ratings for each logo and our explanatory variables are observed at the level of the logo, not the level of the rating. In our baseline model (column 7-1), we find that blind contests result in higher quality entries. This result is marginally significant and in the *opposite* direction of our hypothesis, which predicted that unblind contests would return better entries on average. Recall that there was some evidence of such a relationship, but we believed the balance of evidence would push the net effect in the other direction. Agent talent was a key determinant in that argument, so we attempt to approximate agent expertise and explore this result further.

Although we don't have an independent metric of agent quality or expertise, we approximate it with *MaxI*, a variable that captures the best score an individual has received on prior submissions. If success were random and previous scores weren't predictive of future entries for a given agent, then this metric would be ineffective. However, if we ignore agents who only submit

Table 7. Comparison of Contest Quality between Visibility Treatments

	7-1	7-2	7-3
Dependent variable	Ratings (all entries)	Ratings (all entries)	Ratings (re-submits)
Explanatory variables	Number of entries	Contest results	Contest results
Constant	2.098 ***	2.047 ***	0.428 **
	(0.111)	(0.131)	(0.216)
Treatment			
Blind	0.084 *	-0.318 **	-0.362
	(0.051)	(0.150)	(0.224)
Fixed effects	Yes	Yes	Yes
Explanatory variables			
Entries	0.001	0.001	-0.001
	(0.001)	(0.001)	(0.001)
Entries x Trt[Blind]		0.001	0.007 ***
		(0.002)	(0.002)
Max		0.029	0.105 ***
		(0.022)	(0.026)
Max x Trt[Blind]		0.068	-0.116 **
		(0.048)	(0.045)
MaxI		0.035 *	0.557 ***
		(0.020)	(0.057)
MaxI x Trt[Blind]		0.095 ***	0.237 ***
		(0.034)	(0.069)
R-squared	0.08	0.09	0.20
Mean Response	2.41	2.41	2.40
Observations	11,380	11,380	7,380
DF	45	50	50

OLS regression on individual ratings, clustered by logo, base case is Unblind visibility
Significance levels: * <0.10, ** <0.05, *** < 0.01
Robust clustered standard errors given in parentheses

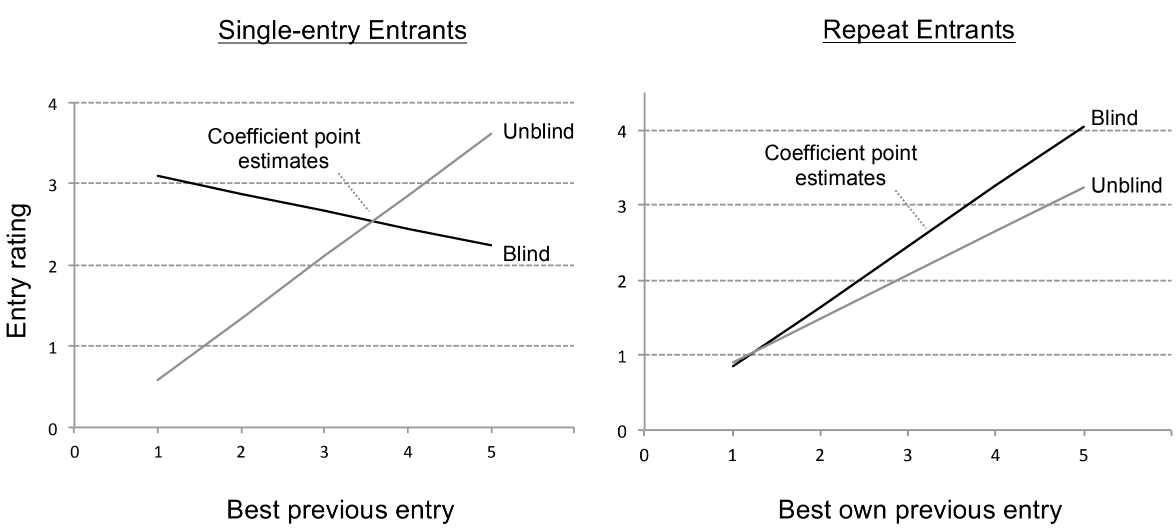
Robust to different measures of agent expertise, including an agent's highest/final MaxI.

once (for which there are no prior scores), the correlation between *MaxI* and rating is 0.30. If we look at an agent's highest *MaxI* globally and compare that talent measure to all their ratings, the correlation is 0.71. In a noisy environment, it indicates there is information in this measure of performance.

Interestingly, when we include our explanatory variables to control for the amount of information in the contest and the performance of the designers, our main effect switches signs (column 7-2). The effect of visibility differs, however, based on agent talent. We observe that unblind contests are better for new entrants (who have no previous best entry) and low-quality designers. High-quality agents perform better in blind contests. Thus, the benefit of entry visibility depends on the type of participants in a given contest.

For this reason, we test one further extension by explicitly modeling just repeat submitters (column 7-3). In this case, low-quality designers in blind and unblind contests submit entries that are identical in quality. As expertise grows, submission quality improves more for blind contests, mirroring the result in column 7-2. If repeating agents are strictly better off in the blind condition, then it is one-time entrants who benefit disproportionately from entry visibility (Figure 3). This lends additional strength to our theory that unblind contests add value by lowering the barriers to entry. Those low-effort designers, who don't submit more than once, benefit from being able to see high quality entries. Looking at *Max* – the best prior entry by others – we see that with

Figure 3. Submission Quality given Search Landscape



Note: Repeat Entrants graph given by coefficient point estimates from model 7-3;
Single-entry Entrants from modified model 7-2 (with only single-entry entrants)

submission visibility, new entries mirror the best existing quality and appear anchored to past results. This effect completely goes away in the blind case, as one would expect. If agents can't see other entries, submission quality decreases with better prior entries, consistent with economic theory around incentive effects (Boudreau et al. 2011). These findings support our hypothesis in part, but also add a new layer of understanding to the tournament literature.

Our final measure of interest is variance in quality. Table 8 starts with a baseline model (column 8-1) that relates our de-trended measure of *quality variance* to *entry visibility* and includes contest fixed effects – *domain* and *site*. We find that increasing visibility (from blind to unblind) reduces the variance in quality we see in the submission ratings. When including *day*

Table 8. Comparison of Contest Variance between Visibility Treatments

	8-1	8-2	8-3
Dependent variable	Variance (de-trended)	Variance (de-trended)	Variance (de-trended)
Explanatory variables	Contest fixed effects	Contest day	Interaction with entries
Constant	0.464 *** (0.034)	0.408 *** (0.047)	0.454 *** (0.049)
Treatment			
Blind	0.090 *** (0.028)	0.088 *** (0.028)	0.023 (0.047)
Fixed effects	Yes	Yes	Yes
Timing			
Day		0.011 * (0.007)	-0.014 (0.011)
Entries			
Entries			0.001 ** (0.001)
Entries x Trt[Blind]			0.003 ** (0.001)
R-squared	0.13	0.13	0.15
Mean Response	0.26	0.26	0.26
Observations	665	665	665
DF	5	6	8

OLS regression on quality de-trended variation at the logo level, adjusted for Judge, base case is Unblind
Significance levels: * <0.10, ** <0.05, *** < 0.01
Standard errors given in parentheses

and *number of entries* as explanatory variables (columns 8-2 and 8-3), our results hold, with variation in the blind setting growing with number of entries. This trend is reasonable; the differences in contest visibility grow over time, as more aggregate information is available. These findings are consistent with Hypothesis 4.

7 Discussion

To understand and characterize the implications of a relatively new decision afforded to innovation contest administrators – that of entry visibility – we examined two primary pathways of influence: (1) the likelihood of entry from an agent and (2) the resulting characteristics of entries in a contest. The related hypotheses we pose in Section 3 are largely supported.

In addressing the first pathway, we find that unblind contests generate more entries; however they do this *not* by inducing more entries from existing agents but by increasing the number of agents that participate. For the second pathway, we examine characteristics for both submission similarity and the quality distribution of entries and find the effect of visibility depends on the setting. Unblind contests encourage submissions that are more similar, mostly in the early stages of the contest. For single-entry participants, entry quality "ratchets up" with the best previous entry if it's visible, while moving in the opposite direction if it's not. However, for invested participants who submit more than once, those with better prior submissions improve more in the absence of entry visibility. Variance in entry quality also improves in the absence of entry visibility.

7.1 Managerial Implications

This research is motivated by the managerial question of whether or not an innovation tournament administrator can improve outcomes based on the moderating decisions within the contest. We found strong evidence to suggest that there are very real differences that result from those decisions. While we cannot extrapolate our results to all innovation contests, understanding the implications of participant entry, idea similarity from search, and contest outcomes should permit managers to more effectively tailor contests for optimal output. Specifically, we uncovered three key decisions contest administrators should manage.

First, managers should be aware that barriers to entry are an important consideration. Unblind contests can attract more entrants, likely because they permit easier search. Casual observers can see exemplars to kick-start their idea development. This doesn't increase the number of entries submitted by each solver, but it does get more solvers in the door.

Second, participant motivation has an effect. The learning environment of unblind contests is better than in blind contests for participants that only submit one entry; seeing a good entry prompts them to come up with a better submission. This is not the case for repeat submitters, who produce better ideas in blind contests. So in an internal company tournament where employees are motivated to participate and likely to submit multiple entries, blind contests may promote better quality (and more varied) ideas. However, in a crowdsourced contest via social media, unblind contests will likely provide better access to landscape exploration and learning and consequently a better result.

Third, entry visibility does impact similarity in designs, but less than we imagined. Unblind contests see a higher level of similarity than blind contests, but the effect quickly goes away. The fear that designers will get stuck in one part of the search landscape does not manifest itself in our data. It appears that participants can use other submissions and create incremental variants that are sufficiently different quickly and efficiently.

7.2 Limitations

Given the fact that we performed this experiment with real designers instead of in a lab, we could not use some potentially interesting designs. The reaction of the same individual under different treatments would be interesting and potentially feasible in a lab study, although in our case, it was not possible.

While the contests were constructed to be nearly identical, we made slight changes in the details in order to avoid detection in the marketplaces. Although we control for contest fixed effects, different challenges could attract fundamentally different types of agents, which could introduce unaccounted for bias into our model.

Beyond this, the backdrop that served for our study deserves some mention. Our setting is nice in that it uses real marketplaces and real designers to test these theories. In addition, logos and graphic design are nice in that the whole idea is represented visually. This may help give insight into more complex domains. However, in graphic design contests (such as those around logos), the effort needed to produce any single idea is relatively small, which could also have implications. Unblind contests may be more acceptable in such situations because the level of investment is minimal. Contests requiring more substantial investment or areas with substantial benefits to intellectual property may not flourish under the same conditions.

7.3 Future Work

As the first to look at the differences of entry visibility on innovation contest outcomes, we have just begun our understanding of this moderating decision. The following questions seem promising for future exploration:

- How does entry visibility apply to different settings? There are plenty of administrator decisions that could improve performance depending on the characteristics of the contest, the solvers, and their interaction with the entry visibility design choice.

- Do different classes of problems behave in the same way? Do algorithmic contests match graphic design contests as related to entry visibility?

- If similarity between ideas does get lost in the unblind case fairly quickly, what density of solvers or entries would be required to again pick up on similarity in ideas? Would a less densely populated ideation landscape change this finding?

• Diverse perspectives are seen as a benefit of open innovation (Jeppesen and Lakhani 2010). Here, we used a pool of solvers from an established contest platform. Controlling for innate solver characteristics would be an interesting direction to further extend the understanding from the level of the agent.

Appendix A. Sample Design Briefs

99Designs – Power Perk Coffee

BUSINESS NAME:
Power Perk Coffee

DESCRIPTION OF BUSINESS:
A better cup of coffee at home – Power Perk focuses on the best ingredients, processes, and accessories for coffee drinkers.

PREFERRED LOGO TYPES:
None specified

COLOR PREFERENCES:
No restrictions on color

TO BE USED ON:
Print (Business cards, letterheads, brochures etc.)
Online (Website, online advertising, banner ads etc.)
Merchandise (Mugs, T-shirts etc.)

NOTES:
Branding - Logo should work across the entire line of coffee products (beans, percolators, and accessories).

Demographics - Our target audience is young adult coffee drinkers (18-35 years old) in the US who are college-educated.

CrowdSpring – Bold Brew Tea

WHAT IS THE EXACT NAME YOU WOULD LIKE IN YOUR LOGO?
Bold Brew Tea

DO YOU HAVE ANY OTHER INFO OR LINKS YOU WANT TO SHARE?
Industry - Home Tea Brewing. Tea leaves, brewing systems, and other accessories for tea drinkers.

Demographics - The focus is on the young adult market in the US. 18-35 year olds who are college-educated and discovering tea as a great beverage alternative.

WHAT ARE THE TOP 3 THINGS YOU'D LIKE TO COMMUNICATE THROUGH YOUR LOGO?
The brand should work over the whole line of tea products. High quality ingredients and processes are the foundation for our image and great-tasting product.

WHAT LOGO STYLES DO YOU LIKE (IMAGE + TEXT, IMAGE ONLY, TEXT ONLY, ETC.)
- Any colors/styles
- No restrictions

Leaps in Innovation:
The Effect of Discontinuous Progress in Algorithmic Tournaments

This paper explores whether innovation breakthroughs stimulate or impede future progress in individual innovation. On the one hand, one could argue that substantial improvements to the status quo might inspire advances through competition. On the other hand, one could claim such improvements might have the opposite effect, stifling motivation or creativity in rivals. Using a unique data set of predictive modeling contests from Kaggle we analyze 25,898 distinct attempts at innovation. We address two related questions to frame our central theme: (1) What effect do discontinuous leaps (as opposed to incremental steps) in innovation contests have on future progress? (2) What predicts such discontinuous leaps in innovation contests? The answers to these questions are as follows. Behavior after discontinuous leaps differs from behavior after continuous steps in innovation tournaments. We find that leaps result in increased rates of entry submission and a speedier turnaround until the next innovation. We also find limited support for leaps improving the trajectory of progress. For the second question, the entrant characteristics that predict leaps turn out to be quite different than those that predict steps. Prior performance, number of prior entries, and platform experience all benefit teams generating incremental improvements. Those same characteristics are not beneficial for innovative breakthrough, mandating a different approach for leaps in innovation. This paper contributes new understanding to the literature on innovation tournaments and offers managers guidance about how to foster leaps in innovation.

1 Introduction

While innovation has long been heralded for its importance in economic development and growth, its ambiguous, uncertain, and fuzzy nature has historically presented sizable obstacles to precise empirical study (Kuznets 1962, Kline and Rosenberg 1986, Fagerberg 2006). Recently, contests have gained renewed popularity as a catalyst for innovation. High-profile challenges, like the $1M Netflix Prize for movie recommendations, have handed important company problems to the public to solve. Not only have innovation tournaments been shown to be effective mechanisms for producing high quality ideas (Terwiesch and Xu 2008, Terwiesch and Ulrich 2009), but recent developments in information technology have resulted in online environments that permit new formats and rapid iteration (Wooten and Ulrich 2013). With electronically captured information related to submissions, participants, and performance, online contests present an opportunity to examine innovation processes under an empirical lens.

Within academic literature, the term innovation has been used in conjunction with many concepts, often with various meanings. Creativity and innovation are often used interchangeably in studies, with innovation tending to include implementation or market factors when a distinction is made (Van de Ven 1986). Definitions of innovation have been used to codify the type of technology advancement, the process by which things are brought to market, the way in which a population perceives a solution, and dynamics within and between firms or organizations. For our setting, we focus on innovation at the level of the individual, with innovation specifically referring to a solution that is better than the best prior solution for a given problem.

Empirically addressing research questions about the evolution of innovation is challenging for a number of reasons. First, there exists a measurement challenge. Objective measures of innovation outcomes are often imprecise or non-existent. Second, there exists a timing challenge. Innovations often go through drastic modifications along the way to being successful (Kline and Rosenberg 1986), so determining defined points in time for measurement is subjective. Third, there exists a comparison challenge. Successive innovations often occur over long timelines, subject to different environments or market conditions. Such factors complicate comparisons both across and within markets. Fourth, there exists a selection challenge. Most solutions aren't realized; without the entire distribution of innovation outcomes, any sample is distorted and captures a view of innovation biased toward the winners.

We develop a unique data set to satisfy the above challenges with contests from Kaggle, an online platform dedicated to data prediction tournaments. Kaggle helps companies publically post their data algorithm problems and generate solutions from data scientists all over the world. As an example, one contest asked participants to predict travel time on Sydney's M4 freeway from past travel time observations. Another sought to improve credit-scoring calculations by predicting the probability that someone experiences financial distress within two years. In all contests, participants get the same starting data and are tested identically (entrants must submit predictions in order to determine how well a particular algorithm scores). Participants are permitted multiple entries over the course of a contest, and we analyze all 25,898 of those solutions across 16 contests.

Our fundamental question looks at the effect of innovation progress on future innovation efforts. Do improvements in innovation stifle additional efforts or inspire further development? To answer this, our paper explores the antecedents and consequences of two types of innovation – continuous (or incremental) and discontinuous (or radical) – in innovation tournaments. Refining or improving in small steps, in line with an existing trajectory, characterizes continuous innovation (Nelson and Winter 1982, Dewar and Dutton 1986, Gatignon et al. 2002), while discontinuous innovation involves disrupting that trajectory (Dosi 1982, Green et al. 1995). Within the contest setting, we refer to specific continuous improvements as *steps* and specific discontinuous improvements as *leaps*. We address two related questions to frame our central theme: (1) What effect do discontinuous leaps (as opposed to incremental steps) in innovation contests have on future progress? (2) What predicts such discontinuous leaps in innovation tournaments?

Our findings are as follows. The answer to our primary question is that behavior after discontinuous leaps in innovation differs from behavior after continuous steps in innovation tournaments. We find that leaps result in an increased rate of entry submission and a speedier time until the next innovation. We also find limited support for leaps resulting in increased average quality after the leap and increased rate of contest trajectory late in contests. For our second question, the entrant characteristics that predict leaps turn out to be quite different than those that predict steps. Prior performance, number of prior entries, and platform experience all benefit teams that are refining iteratively and ultimately generate an incremental improvement. Those same beneficial characteristics do not play a role in leaps of innovation; in fact, having a higher prior score works against an innovative breakthrough.

2 Literature and Hypothesis Development

In this paper, we examine cause and effect relationships of innovation progress in contests. Contests have been shown to be effective mechanisms for generating new solutions (Terwiesch and Xu 2008, Terwiesch and Ulrich 2009). The problem of underinvestment that comes with increased participation is offset by the benefit of lots of potential solvers and parallel search (Terwiesch and Xu 2008, Boudreau et al. 2011). Such results have focused on blind, one-shot innovation tournaments. However, new types of contests are gaining both feasibility and popularity thank to recent developments in information technology – including unblind contests (Bockstedt et al. 2012, Wooten and Ulrich 2012) and sequentially scored contests with repeated entry (Wooten and Ulrich 2013). If tournaments have repeated entry, then participants can submit multiple times, learning with each iteration. In some cases, performance can be determined and displayed in real-time, and often, public leaderboards show the scores of all contestants. Such setups result in more dynamic environments, with elements of learning and competition naturally embedded. In many ways, repeated entry contests mimic a traditional innovation setting more than one-shot contests. We focus on the dynamic progress that occurs over the course of contests with repeated entry and consider the impact of innovation improvements on the contest outcome. We review the literature on competition, incentives, and creativity to guide us as we develop our hypotheses.

2.1 Competition and Incentives

In what ways might a discontinuous leap in innovation impact tournament outcomes? Increases in the best solution could stifle additional efforts or inspire further development. On the one hand, one could argue that substantial improvements to the status quo might inspire advances through competition. On the other hand, one could claim such improvements might have the opposite effect, stifling motivation or creativity in rivals.

Holding everything else equal, improvement in an innovation tournament does two things: it pushes the innovation frontier out (raising the bar on what is innovative) and it signals what is possible (which in solving unknown problems is not insignificant). These two forces work in opposition, with the former moving the goal farther away and the latter bringing a target into the realm of possibility.

Competition can impact both of the above pathways. While competition has long been stressed as a positive influence on innovation, especially within the economics literature (Greenhalgh and Rogers 2006), it has enjoyed a somewhat more ambiguous relationship in the creativity and management literature (Li and Vanhaverbeke 2009). Bullinger et al. (2010) give a nice summary of those two streams of research, with the conclusion that innovation contests offer a competitive setting where participants feel challenged and unconstrained, which leads to a high degree of creativity and innovativeness. This meshes with the findings that a controlling environment is detrimental to creativity and that the effect of competition may be negative within an organization, but not between organizations (Amabile 1996).

Anecdotally, there is some evidence that competition in this innovation tournaments spurs additional effort and creativity. Boudreau et al. (2011) observe that entrants (especially those in the higher-performing echelons) take it personally if they don't end up on top in tournaments on TopCoder, a software contest platform. A scientist at IBM Research (and top performer in one of the Kaggle contests) echoed these thoughts when asked about outcomes. He emphatically stated that just given a dataset, he would not have performed nearly as well – that he was motivated to try new things after others overtook him on the leaderboard. These give some indication of the motivating power of competition.

However, such effects are not uniform. In a study of rank-order algorithm tournaments, average performance decreased when more superstars participated; a small group of high-ability individuals just below the superstars instead reacted positively to the increased competition, with increased effort but without additional errors of logic (Boudreau et al. 2012). In another study of sales contests, effort put forth by participants trailing the leader only faded when the gap was very large, but front-runners reduced their effort as their lead extended (Casas-Arce and Martínez-Jerez 2009).

In addition to effort, creativity benefits from increasing competition, but only up to an intermediate level of intensity (Baer et al. 2010). In online platforms, where the presence of the competition may be less salient, overall creativity is likely to be higher with competition than without (Shalley and Oldham 1997). We conjecture that contest participants perceive innovation leaps – below a certain threshold – as greater competition than incremental steps. Combining these prior results, we predict that leaps in innovation will lead to increased effort and creativity in contests when compared to continuous steps, despite the greater gap that must be overcome.

Hypothesis 1: Discontinuous leaps in innovation contests will increase the effort of entrants, resulting in more entries per time period.

Hypothesis 2: Discontinuous leaps in innovation contests will increase the performance of entrants, resulting in an improved contest outcome.

2.2 Creativity

Given that we expect greater effort and performance, a natural next question is how to predict (and encourage) leaps in the first place. The components of individual and team creativity that lead to innovation are typically framed as expertise, creative thinking, and task motivation (Amabile 1996). Expertise is defined here as foundational know-how, like factual knowledge or technical proficiency. Creative thinking is a set of skills – like cognitive openness to new perspectives – that helps explore new pathways. Task motivation includes both intrinsic and extrinsic varieties as well as a baseline component and variable component at any given point in time.

One predictor of future ideation outcomes is past experience. Experimental research shows that cognitive fixation – where past experiences impede future efforts – is widespread in creativity exercises, including creative idea generation (Smith 2003). A similar idea suggests that creative ideation thrives under non-normal conditions and that constraints and guidance work to creativity's detriment (Csikszentmihali 1996). Bayus (2013) gives a nice summary of the literature in creativity and cognitive psychology that highlights these relationships and goes on to find that a successful idea from a serial ideator has a detrimental effect on finding another winner – serial ideators are more likely than single-submission entrants to have one success, but not multiple. The opposite has been championed as well. Some individuals may simply be better at generating high-quality ideas (Terwiesch and Ulrich 2009), which suggests that past success should be positively correlated to future success. In a series of unblind graphic design contest experiments, there was a significant, positive association between an entrant's best prior entry (as well as the best prior entry by others) and their future entries (Wooten and Ulrich 2012).

Other work with innovation tournaments suggests who might be more likely to achieve problem-solving success. InnoCentive's R&D tournaments see increased likelihood of success for solver's whose own field of expertise is farther away, in terms of technical similarity, from the problem domain (Jeppesen and Lakhani 2010). Natural experiments on LogoMyWay.com show that entrants more likely to win join earlier and have a wider range of entry timing; however, simply increasing the number of entries offers no benefit (Bockstedt et al. 2012). In the same study, prior experience – and especially winning experience – increases the likelihood of winning.

We consider how team characteristics impact each of our creativity attributes. First, we expect expertise to be positively associated with team experience and – to a lesser degree – number of members on a team. Task motivation is likely to be impacted by a team's best prior score and number of entries. Teams are likely to be more motivated the closer they are to winning, and number of entries should both act as a signal of motivation and also reinforce that tendency. Creative thinking is less straightforward, but based on the cognitive fixation work, we think that a team's best prior score could work in opposition to creativity. In our data science competitions, we conjecture that being near the top of the leaderboard is motivating but also limits the way in which an entrant searches for their next algorithm. We would expect that achieving a good score would results in a local maxima strategy of search, whereas teams farther down the leaderboard might display an "openness to new perspectives" and search farther afield. Thus, our characteristics all load the creativity attributes in the same direction except for a team's best prior score. We believe a creative search strategy may be more likely to produce a discontinuous leap but recognize that it could go either way. We summarize these as the following testable hypotheses.

Hypothesis 3: The likelihood of a discontinuous leap in an innovation tournament is *decreasing* in an entrant's best prior score and *increasing* in number of entries from an entrant, number of team members, and experience.

Hypothesis 4: The likelihood of a continuous step in an innovation tournament is *increasing* in an entrant's best prior score, number of entries from an entrant, number of team members, and experience.

3 Data and Methods

To address the hypotheses posed above, we rely on detailed contest data from an algorithm tournament platform. Our aim is to analyze the effect of innovation progress on future innovation efforts, which we do by examining individual solutions and the overall trajectory within the contest framework. Do improvements in innovation stifle additional efforts or inspire further development? Our setting is also useful in addressing another question that naturally surfaces in the course of examining the differences between leaps and steps in innovation: What predicts discontinuous leaps (versus incremental steps) in innovation tournaments?

Empirically addressing research questions about the evolution of innovation is challenging for a number of reasons. First, there exists a measurement challenge. Objective measures of innovation outcomes are often imprecise or non-existent. Second, there exists a timing challenge. Innovations often go through drastic modifications along the way to being successful (Kline and Rosenberg 1986), so determining defined points in time for measurement is subjective. Third, there exists a comparison challenge. Successive innovations often occur over long timelines, subject to different environments or market conditions. Such factors complicate comparisons both across and within markets. Fourth, there exists a selection challenge. Most solutions aren't realized; without the entire distribution of innovation outcomes, any sample is distorted and captures a view of innovation biased toward the winners.

Our setting is one in which observational data presents itself as a quasi-experiment in a real-world setting. With quantifiable outcomes thanks to their data-driven nature, algorithm contests are well suited for examining innovation. Within each contest, teams operate independently and we treat discontinuous leaps as exogenous disruptions to the system.[1]

3.1 Contest Platform

Kaggle, the leading online platform in the crowd-sourced predictive modeling market, provided the data we analyze. Since 2010, Kaggle has hosted public data science competitions, where sponsors post their problems and data scientists from all over the world compete to create the best solution. As an example, one contest asked participants to predict travel time on Sydney's M4 freeway from past travel time observations. Another sought to improve credit-scoring calculations by predicting the probability that someone experiences financial distress within two years. Companies such as General Electric, Merck, Allstate, Facebook, and Ford have participated.

Contest winners are awarded predetermined cash prizes whose amounts vary significantly, with most between $1,000 and $100,000. Typical contests last 2-4 months. In its first three years, Kaggle has hosted 85 public competitions (as well as 100 private competitions for school classroom use) for its community of over 83,000 data scientists, amounting to more than $4M in cash prizes. Our sample of contests is restricted to cash prize contests from its first two years of operation.

As Boudreau et al. (2011) describe, participant motivation is a central concern in innovation tournaments. In addition to the contest awards, Kaggle also has a for-hire practice that rates and matches talent from its contests with companies looking for project consultants. Facebook has

[1] Future robustness checks could include exploring ways to identify and correct unobserved heterogeneity and endogeneity within the model, including running first stage estimators as predictors.

[2] Contest fixed effects can also be modeled as categorical variables for each contest, which the model is

even run two contests where the winning prize was a job at Facebook, so there is real value in the platform for its participants.

3.2 Algorithm Contest Details

At any given time, a dozen or so contests (Figure 1) are active on Kaggle. Each contest includes a competition overview, data sets, a leaderboard, and administrative details (including the prize amount, deadlines, and number of participating teams). The leaderboard is visible to the public and shows how well each of the participating teams is doing in the competition. It contains each team's name, member profiles, current rank, one-week change in rank, score, number of entries, and timestamps best and last submissions. Participants do not observe the actual submissions of entrants, only the informational data above. A sample leaderboard is shown in Appendix A.

Once a contest has begun, participants may download the data sets and begin submitting entries to Kaggle for scoring. There are typically two distinct data sets. The first is a training set, which is used to develop an algorithm and includes filled-in values for the response variable of interest. The second set is the test file and has the response variable omitted. An entry consists of a team's predictions for the missing response variables in the test file. Winners are chosen based on the ability to accurately predict the missing responses. Kaggle scores entries by comparing teams' predictions to the actual answers. A fraction of the test data set (usually 25-33%) provides scores for the public leaderboard (Appendix A). Final placement is calculated off of the reserved, private portion of the test set. There is a cap on the number of submissions per day – typically 1 or 2. Teams of more than one person are allowed.

3.3 Data and Variables

Kaggle granted us access to the full database records of their contests over a period from 2010-2011, which included 26,082 entries in 23 contests. After removing those contests that did

Figure 1. Sample Kaggle Contest Page

kaggle

Customer Solutions ▾ | Competitions | Community ▾ | Sign Up | Login

MARINEXPLORE WHALE DETECTION

The Marinexplore and Cornell University Whale Detection Challenge

2.7 days to go

Friday, February 8, 2013 — $10,000 • 223 teams — Monday, April 8, 2013

Dashboard

- Home
  - Information ▸
  - Data
  - Make a submission
- Forum
- Leaderboard
- Visualization

Leaderboard

1. SluiceBox (60)
2. alfnie (15)
3. Free Willzyx (32)
4. Jure Zbontar (24)
5. Daniel Nouri (16)
6. Tree growers (77)
7. RBM (24)
8. John Clarke (34)
9. megasoft (18)
10. Nico de Vos (14)

Forum (28 topics)

Quiet periods / Serial correlation
yesterday

Features & classification approaches
2 days ago

Repeat waveforms
9 days ago

To read aiff format using Perl
10 days ago

Anyone looking to form a Team?
12 days ago

Software to be used?
17 days ago

223 teams with

283 participants

2827 entries

Competition Details » Get the Data » Make a submission

Create an algorithm to detect North Atlantic right whale calls from audio recordings, prevent collisions with shipping traffic

We depend on shipping industry's uninterrupted ability to transport goods across long distances. Navigation technologies combine accurate position and environmental data to calculate optimal transport routes. Accounting for and reducing the impact of commercial shipping on the ocean's environment, while achieving commercial sustainability, is of increasing importance, especially as it relates to the influence of cumulative noise "footprints" on the great whales.

Marinexplore is organizing the Planet's ocean data with the leading community of ocean professionals. One of the important datasets consists of acoustic recordings that can be used to detect species inhabiting the global ocean. Knowledge about animal locations can be utilized in industrial operations.

Cornell University's Bioacoustic Research Program has extensive experience in identifying endangered whale species and has deployed a 24/7 buoy network to guide ships from colliding with the world's last 400 North Atlantic right whales.

Cornell University MARINEXPLORE

Illustration of ships navigating safely around the habitat of whales.

Right whales make a half-dozen types of sounds, but the characteristic up-call is the one identified by the auto-detection buoys. The up-call is useful because it's distinctive and right whales give it often. A type of "contact call," the up-call is a little like small talk--the sound of a right whale going about its day and letting others know it's nearby. In this recording the up-call is easy to hear--a deep, rising "whoop" that lasts about a second:

Marinexplore and Cornell researchers challenge YOU to beat the existing whale detection algorithm identifying the right whale calls. This will advance ship routing decisions in the region.

[For details on the buoy network see a paper published by Acoustical Society of America.]

Started: **12:00 am, Friday 8 February 2013 UTC**
Ends: **12:00 am, Monday 8 April 2013 UTC (59 total days)**

not allow repeated entry, were publicity stunts, or were test launches, our sample consisted of 25,898 entries across 16 contests. As these are blind contests with repeated entry, entrant information is restricted to their own private information and that which is publically available on the leaderboard. To measure effort, performance, and team characteristics, we observe a set of contest variables (Table 1).

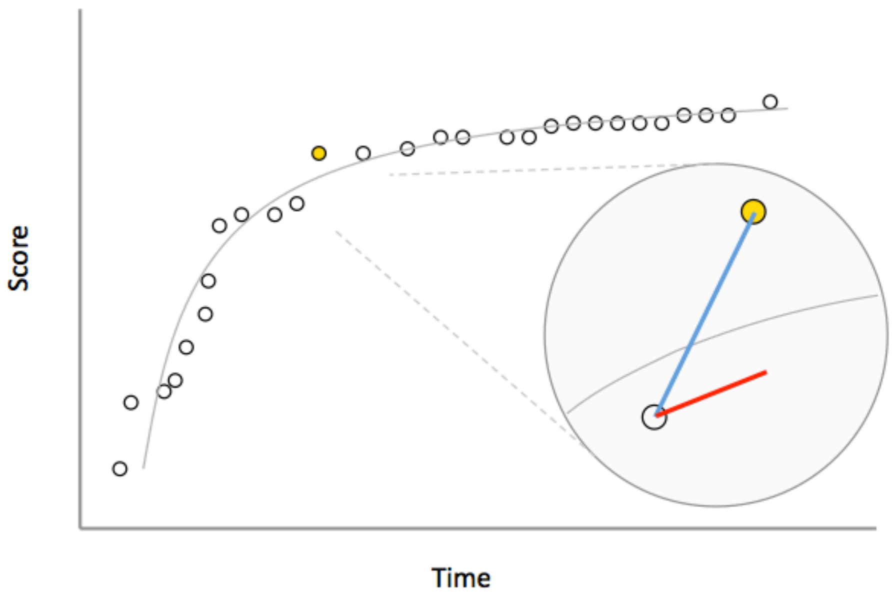
3.4 Measuring Discontinuous Leaps in Innovation

In order to analyze leaps in innovation, we must be able to measure them. Here, we outline our procedure for identifying and which improvements in a contest count as steps and which count as leaps. This is complicated by the fact that ideation and trial-and-error projects display diminishing improvements over time, with each increase requiring more entries in expectation (Terwiesch and Xu 2008, Terwiesch and Ulrich 2009). Thus, with something approximating a logarithmic return function, we take into account the fact that the trend over time is dynamic. Evaluating the contests ex post, we fit each contest with a logarithmic trend line. Any contest maximum more than 5% above the point that the trend line would predict (calculated from the

Table 1. Data Definitions

	Variable	Definition
(1)	Leap	Indicator of discontinuous jump; 1 if entry improves max 5% more than expected
(2)	Step	Indicator of continuous jump; 1 if improves the contest max and is not a leap
(3)	Progress	Elapsed number of days/number of days contest ran
(4)	Day	Control for day of the week - 1: Sun, 2: Mon, … 7: Sat
(5)	Entries	Number of prior entries in contest at time of submission
(6)	Score	The score awarded to a given entry
(7)	Team's best prior	Best prior score awarded to a given entrant
(8)	Team's # of entries	Number of prior entries by an entrant at time of submission
(9)	Team's # of members	Number of individuals on a team
(10)	Team experience	Indicator of participation in a prior Kaggle contest; 1 if experienced
(11)	Prize amount	Control for cash award in contest
(12)	Contest length	Control for number of days contest ran
(13)	Contest appeal	Control for number of unqiue entrants; standardized as entrants/day

Figure 2. Identification of Innovation Leaps



previous best point) is flagged as a leap. In Figure 2, the highlighted entry is flagged because the distance between the blue and red endpoints on the right exceeds our 5% threshold. This methodology flags 37 leaps over 16 contests.

3.5 Evaluation

Table 2 contains the descriptive statistics on the variables. Because of the diminishing rate of improvement in contests, differences or difference-in-difference measures are used for before-after comparisons. For the first two hypotheses, we capture *change in entry velocity* to measure effort and *change in average score*, *change in adjusted rate*, and *# entries until next step* to measure performance.

Change in entry velocity is defined as the number of entries appearing in the 5% of contest time after a particular entry minus the number of entries appearing in the 5% of contest time prior. Entries from the entrant are excluded from both segments.

Change in average score is the average of the 15 next unique submissions minus the average of the 15 prior unique submissions.

Change in adjusted rate is defined as the difference between the slope-adjusted rate of improvement of the max score after a particular entry minus the same improvement rate before, over a 30-entry window each direction.

Number of entries until next step is the count of contest entries until another unique improvement occurs.

Table 2. Descriptive Statistics

		Entry-level		*Contest-level*	
	Variable	Mean	SD	Mean	SD
(1)	Leap	0.001	0.037	0.005	0.008
(2)	Step	0.015	0.121	0.025	0.023
(3)	Progress	0.585	0.303	0.598	0.071
(4)	Day	4.021	1.946	4.032	0.233
(5)	Entries	1304.616	985.880	809.019	654.120
(6)	Score	0.776	0.235	0.692	0.209
(7)	Team's best prior	0.763	0.300	0.671	0.198
(8)	Team's # of entries	16.253	20.884	13.456	6.858
(9)	Team's # of members	1.106	0.438	1.114	0.169
(10)	Team experience	0.326	0.469	0.382	0.192
(11)	Prize amount			0.190	0.749
(12)	Contest length			71.097	28.999
(13)	Contest appeal			2.747	2.158

4 Results

4.1 Effects of Leaps in Innovation

We answer our main question about the effect of leaps in innovation on future innovation outcomes by estimating relationships with our four dependent variables of interest:

change in entry velocity	how frequently were entries submitted,
change in average score	how good was the population of adjacent entries,
change in adjusted rate	how quickly was the best entry improving,
# entries until next step	how long until another improvement of any kind.

The first three of these are difference measures, as discussed in Section 3.5, to account for the dynamic nature of the contests. To control for differences across contests, which could influence our behavior measures if not accounted for, we include several fixed effects – *prize amount*, *contest length*, and *contest appeal* – in all models.[2] Additionally, we use a randomized 5% subset of data points that exhibit no improvement to include with those entries that did improve the contest max.[3] Table 2 provides variable details and gives descriptive statistics and correlations for the variables used in our analysis.

Table 3 shows the results of our linear regressions for the above four dependent variables. The analysis begins by looking first at the effect that *leaps* and *steps* have on *change in entry velocity* (column 3-1 and 3-2). We include controls for day of the week since fewer entries are submitted on the weekend. This shows up as negative coefficients on Wednesday through Saturday because our dependent measure extends several days in each direction from the day of measurement. So Friday measurements pick up the entry velocity over the weekend and then subtract out the entry velocity from the week prior; the daily patterns are accounted for here, giving better predictions for the model. In both cases, the main effects show that leaps in innovation lead to higher rates of entries immediately afterward. The coefficients for step are significant and smaller than those for leap, in line with our first hypothesis (H1). One interpretation is that seeing a big improvement on the contest leaderboard carries significantly more motivational weight than an incremental improvement, with teams putting forth more effort as a result.

The next question is whether that effort translates to better contest entries. Our next three models develop that analysis. One measure of improvement is *change in average score* (column 3-3). If entry scores improve on average after new contest bests, then the extra effort we observed

---

[2] Contest fixed effects can also be modeled as categorical variables for each contest, which the model is robust to. By using the characteristics instead, we gain additional insight into the tournament behavior.

[3] We do this to avoid over-sampling. Each of our dependent measures includes entries before and after it; if we included every point, we'd be including the same values multiple times over in various difference-in-difference measurements.

Table 3. Comparison of Leaps and Steps on Subsequent Innovation

	3-1	3-2	3-3	3-4	3-5
Dependent variable	Change in entry velocity	Change in entry velocity	Change in avg score	Change in adjusted rate	# Entries until next step
Explanatory variables	Contest fixed effects	Progress controls	Progress controls	Progress controls	Progress controls
Constant	9.597	-20.746 ***	0.038 *	-0.051 ***	-73.742 *
	(7.564)	(7.786)	(0.020)	(0.007)	(45.864)
Explanatory variables					
Leap	-1.965	30.934 **	0.143 ***	-0.134 ***	53.421
	(10.252)	(13.516)	(0.043)	(0.022)	(133.931)
Step	-6.658 **	18.069 ***	-0.000	-0.002	155.001 ***
	(2.826)	(5.773)	(0.014)	(0.005)	(31.804)
Monday	0.632	1.627			
Tuesday	6.543	4.693			
Wednesday	-12.359 ***	-12.976 ***			
Thursday	-21.642 ***	-23.167 ***			
Friday	-25.995 ***	-27.916 ***			
Saturday	-17.090 ***	-18.578 ***			
Contest progress		51.481 ***	-0.010	0.009 **	26.380
		(4.748)	(0.024)	(0.004)	(26.443)
Contest progr. x Leap		-65.516 *	-0.189 **	0.190 ***	-74.357
		(36.524)	(0.076)	(0.033)	(192.486)
Contest progr. x Step		-42.550 ***	0.001	0.003	-13.595
		(9.951)	(0.024)	(0.009)	(54.039)
Control variables					
Prize amount	-10.018 ***	-9.740 ***	0.000	-0.005 ***	17.251
	(1.926)	(1.850)	(0.005)	(0.002)	(10.904)
Contest length	0.012	0.027	-0.000 **	0.000 ***	1.390 ***
	(0.064)	(0.062)	(0.000)	(0.000)	(0.399)
Contest appeal	3.684 ***	4.041 ***	-0.001	0.005 ***	23.345 ***
	(0.934)	(0.897)	(0.002)	(0.001)	(5.336)
R-squared	0.10	0.17	0.01	0.06	0.01
Mean response	8.3	8.3	0.0	-0.0	163.3
Observations	1,341	1,341	1,572	1,522	1,572
DF	11	14	8	8	8

OLS regression, standard errors given in parentheses
Significance levels: * <0.10, ** <0.05, *** < 0.01

is translating to results. This measure looks at the 15 scores before and after the entry in question. The coefficient observed for leap is positive and significant, compared to both no improvement and step improvement. Looking at the interaction effect, however, we note that this advantage goes away by the end of the contest. There isn't much signal here, even given the support for Hypothesis 2.

Our second test for evaluating if entries become more innovative looks at the rate of contest improvement before and after a specific entry (column 3-4). Since the rate of improvement diminishes over the course of the contest, this is an adjusted measure that also factors in that expected decay, as discussed in Section 3.5. While our main effect shows a significantly negative value – meaning that the rate of improvement after leaps isn't as steep as the rate of improvement otherwise – factoring in the progress of the contest, we observe that after 70% of a contest has elapsed, leaps result in higher growth.

Our last regression around outcomes relates the *number of entries until next step* to whether or not leaps or steps were present (column 3-5). Here, we observe that after a continuous step improvement, it takes significantly more entries to witness another innovation improvement than it does after a discontinuous leap. Thus, between our three measures of outcomes, we observe some differences favoring leaps. The implication is that since leaps should be more difficult to overcome than steps, that the presence of a leap is spurring additional task motivation and/or additional creative thinking.

4.2 Creating Leaps in Innovation

Having determined that leaps in innovation can actually inspire better performance, a natural follow-up question is how to encourage leaps in the first place. Understanding what predicts discontinuous leaps in innovation tournaments thus becomes the second question within our unique setting.

Table 4 shows the results of a logit regression analysis around predicting innovation improvements. We begin by estimating the baseline models (column 4-1 and 4-3) by relating the likelihood of an innovation occurring for a given entry to that *team's best prior* entry. We again include our contest fixed effects.

We find that prior performance is significant in both cases but acts in opposite ways for the two types of innovation improvement. The likelihood of generating an innovation leap is negatively associated with a team's best prior entry (and a step improvement is positively associated with a team's best prior entry). This suggests that continuous step improvements originate from those entrants who have previously performed well, likely through a refinement of that prior solution. This is consistent with the idea of local maxima and steady improvement. Leaps, however, come from those whose prior ideas were not as successful. The idea of cognitive fixation (Smith 2003) plays out in this dynamic, with well-performing groups not able to create a leap.

Table 4. Prediction of Leaps and Steps in Innovation

	X-1	X-2	X-3	X-4
Dependent variable	Leap	Leap	Step	Step
Explanatory variables	Contest fixed effects	Entry descriptors	Contest fixed effects	Entry descriptors
Constant	-2.588 ***	-0.952	-3.638 ***	-3.069 ***
	(0.631)	(0.882)	(0.299)	(0.356)
Explanatory variables				
Team's best prior	-2.227 ***	-1.035 **	1.203 ***	1.062 ***
	(0.420)	(0.523)	(0.230)	(0.241)
Team's # of entries		-0.032		0.021 ***
		(0.032)		(0.002)
Team's # of members		0.226		0.489 ***
		(0.319)		(0.064)
Team experience		-0.340		0.243 **
		(0.382)		(0.115)
Contest progress		-4.582 ***		-2.754 ***
		(0.913)		(0.222)
Control variables				
Prize amount	0.949 **	0.934 **	0.107	0.125
	(0.366)	(0.378)	(0.107)	(0.112)
Contest length	-0.015 **	-0.019 **	-0.008 ***	-0.009 ***
	(0.007)	(0.008)	(0.003)	(0.003)
Contest appeal	-0.585 ***	-0.652 ***	-0.230 ***	-0.315 ***
	(0.149)	(0.159)	(0.041)	(0.044)
Log-likelihood (-)	242	217	1940	1813
Chi-squared	61 ***	111 ***	101 ***	355 ***
Observations	25,898	25,898	25,898	25,898
DF	4	8	4	8

Logit regression, standard errors given in parentheses
Significance levels: * <0.10, ** <0.05, *** < 0.01

Because our contests occur over time, we extend the model to include *progress* as an explanatory variable; we also include additional team characteristics to capture variation in experience and expertise (column 4-2 and 4-4). The extended results mirror our baseline models. We also see that not only does a team's best prior entry work differently for teams who innovate stepwise versus those who leap, but prior number of entries, number of team members, and team experience are significant for steps but not for leaps. In this setting, creating a leap in innovation wasn't predicated on entering a lot of times, having a big team, or having prior contest experience. The coefficient for *progress* is negative for both leaps and steps, indicating that contests are most likely to see innovation improvements at the beginning of the contest, as one would expect. Both progress and team's best prior are scaled 0-1 variables, making comparisons straightforward. The timing of the contest matters about twice as much for both leaps and steps. Together, this supports Hypothesis 4 and partially supports Hypothesis 3. Only prior score showed up as significant for leaps from our four predictions, although leap also appears to load on *prize*. That suggests that you can pay for leaps, most likely by attracting better talent to the tournament.

5 Discussion

The bulk of this paper deals with distinguishing between two types of improvements that can happen in innovation – the iterative, incremental improvement and the radical, discontinuous leap – and their effect on further innovation.

Behavior after discontinuous leaps differs from behavior after continuous steps in innovation tournaments. We find that leaps result in increased rates of entry submission and a speedier turnaround until the next innovation. Also, the entrant characteristics that predict leaps turn out to be quite different than those that predict steps. Prior performance, number of prior entries, and platform experience all benefit teams generating incremental improvements. Those same characteristics are not beneficial for innovative breakthrough, mandating a different approach for leaps in innovation. Once an innovator is on a well-performing path, they are likely to continue to push the innovation frontier in a continuous manner. This paper contributes new understanding to the literature on innovation tournaments.

5.1 Managerial Implications

First, managers should be aware that contest participants are engaged and pay attention. Even in blind contests, where solutions aren't revealed to the public, participants are affected by the contest, especially through motivation and creativity.

Second, from this exploration (and in contrast to some other tournament settings), it appears that the agents who delivered discontinuous leaps in the contests were not those with high prior scores. This suggests that if you can keep the entire population of solvers active and interested, you won't inadvertently lose a future leaper.

Also, it may be beneficial to seed innovation efforts with leaps. If the setting is conducive to competition, fabricating improvements for the public leaderboard could result in improved contest outcomes. Finally, it might also be possible to purchase leaps in innovation. It appears from our results that higher prize amounts attracted more leaps. We need to investigate this phenomenon further, but another tactic is spending more money to insure that a certain level of innovation participates. Instead of paying for a particular person or agency or firm, extra dollars in a contest means you are buying extra crowd share, which may prove to be an efficient way to drive leaps in innovation.

5.2 Limitations

The use of real-world contests is both a benefit and a hurdle. By examining challenging, reasonably complex problems that real companies seek to solve through tournaments, we avoid many issues of irrelevance. The drawbacks, however, are the challenges encountered without the randomness of a true intervention. Internal validity and confounding variables are a concern, as the real, public nature of these contests requires that lots of elements be outside the control of the investigator.

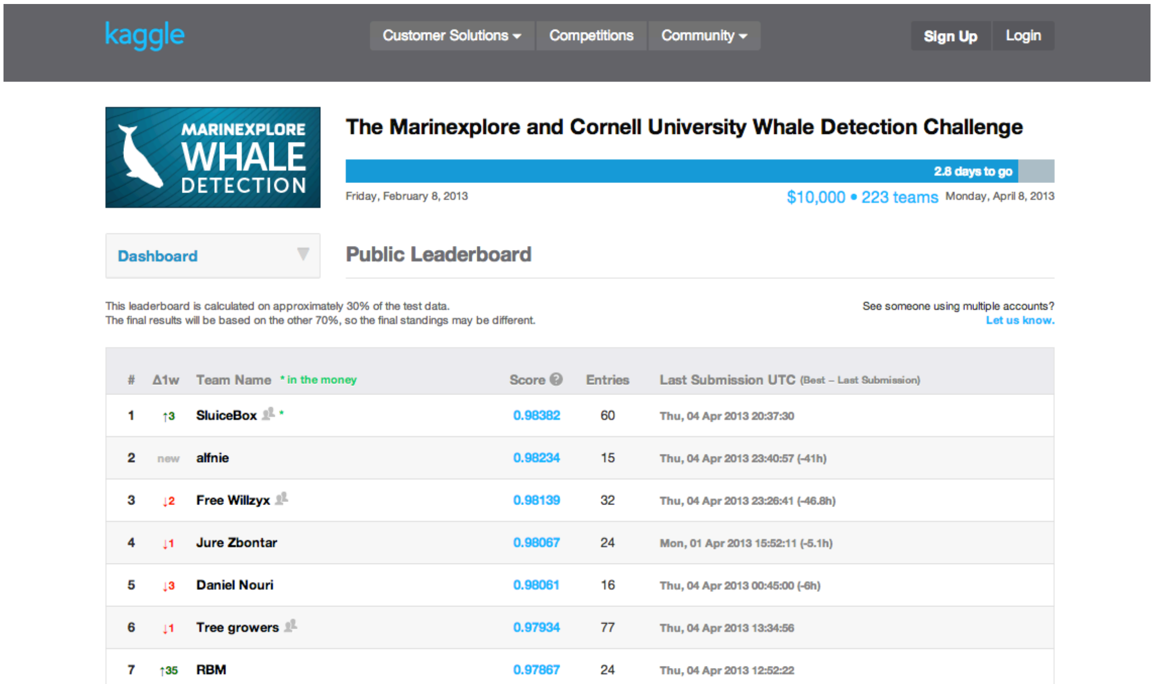
Additionally, we explore one particular type of innovation tournament in this paper. Algorithmic data prediction contests represent only a small fraction of the total contests and innovation efforts being pursued. While we think that the rigorous and challenging nature of these uncertain scientific problems represent general areas of innovation fairly well, the results may not generalize to all blind, repeated-entry tournaments.

5.3 Future Work

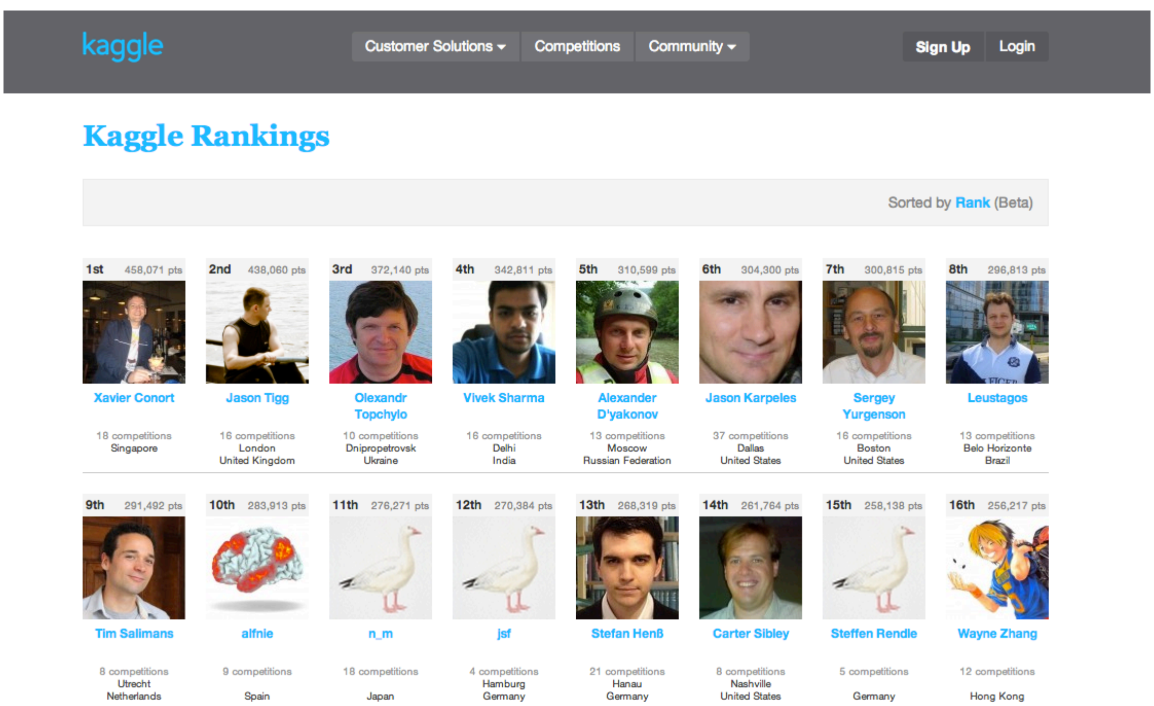
We can imagine other streams of work that rely on innovation tournaments and further explore these questions. The following avenues seem promising:

- How does a conditional view of innovation progress marry with what we know about search in innovation?
- Do different classes of problems behave in the same way? Does uncertainty influence outcomes in new ways?
- Coordinating contest with a platform provider and including artificial *leaps* in a true field experiment setting would lend additional weight to the conclusions of this study.
- Can certain characteristics of leaps be purchased? Is it more expensive to buy expertise or motivation or creativity for your innovation needs?

Appendix A. Sample Kaggle Leaderboard and Site Rankings



#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑3	SluiceBox *	0.98382	60	Thu, 04 Apr 2013 20:37:30
2	new	alfnie	0.98234	15	Thu, 04 Apr 2013 23:40:57 (-41h)
3	↓2	Free Willzyx	0.98139	32	Thu, 04 Apr 2013 23:26:41 (-46.8h)
4	↓1	Jure Zbontar	0.98067	24	Mon, 01 Apr 2013 15:52:11 (-5.1h)
5	↓3	Daniel Nouri	0.98061	16	Thu, 04 Apr 2013 00:45:00 (-6h)
6	↓1	Tree growers	0.97934	77	Thu, 04 Apr 2013 13:34:56
7	↑35	RBM	0.97867	24	Thu, 04 Apr 2013 12:52:22



BIBLIOGRAPHY

ALICKE, M. D., M. L. KLOTZ, D. L. BREITENBECHER, T. J. YURAK, and D. S. VREDENBURG (1995): "Personal Contact, Individuation, and the Better-than-Average Effect," Journal of Personality and Social Psychology, 68, 804-825.

AMABILE, T. M. (1996): *Creativity and Innovation in Organizations*, Harvard Business School.

ARCHAK, N., and A. SUNDARARAJAN (2009): "Optimal Design of Crowdsourcing Contests," ICIS 2009 Proceedings, paper 200.

BAER, M., R. T. A. LEENDERS, G. R. OLDHAM, and A. K. VADERA (2010): "Win or Lose the Battle for Creativity: The Power and Perils of Intergroup Competition," *Academy of Management Journal*, 53, 827-845.

BANDIERA, O., V. LARCINESE, and I. RASUL (2009): "Blissful Ignorance? Evidence from a Natural Experiment on the Effect of Individual Feedback on Performance," working paper.

BARANKAY, I. (2011): "Rankings and Social Tournaments: Evidence from a Crowd-Sourcing Experiment," working paper.

BAYUS, B. L. (2013): "Crowdsourcing New Product Ideas over Time: An Analysis of the Dell IdeaStorm Community," *Management Science*, 59, 226-244.

BOCKSTEDT, J., A. MISHRA, and C. DRUEHL (2012): "Do Participation Strategy and Experience Impact the Likelihood of Winning in Unblind Innovation Contests?," working paper.

BOUDREAU, K. J., C. HELFAT, K. LAKHANI, and M. MENIETTI (2012): "Field Evidence on Individual Behavior and Performance in Rank-Order Tournaments," working paper.

BOUDREAU, K. J., N. LACETERA, and K. R. LAKHANI (2011): "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis," *Management Science*, 57, 843−863.

BULL, C., A. SCHOTTER, and K. WEIGELT (1987): "Tournaments and Piece Rates: An Experimental Study," *Journal of Political Economy*, 95, 1-33.

BULLINGER, A. C., A. NEYER, M. RASS, and K. M. MOESLEIN (2010): "Community-Based Innovation Contests: Where Competition Meets Cooperation," *Creativity and Innovation Management*, 19, 290-303.

CASAS-ARCE, P., and F. A. MARTÍNEZ-JEREZ (2009): "Relative Performance Compensation, Contests, and Dynamic Incentives," *Management Science*, 55, 1306-1320.

CHE, Y., and I. GALE (2003): "Optimal Design of Research Contests," *The American Economic Review*, 93, 646-671.

CSIKSZENTMIHALY, M. (1996): *The Creative Personality*.

DAHAN, E., and H. MENDELSON (2001): "An Extreme Value Model of Concept Testing," *Management Science*, 47, 102-116.

DAHL, D. W., and P. MOREAU (2002): "The Influence and Value of Analogical Thinking During New Product Ideation," *Journal of Marketing Research*, 39, 47-60.

DEWAR, R. D., and J. E. DUTTON (1986): "The Adoption of Radical and Incremental Innovations: An Empirical Analysis," *Management Science*, 32, 1422-1433.

DOSI, G. (1982): "Technological Paradigms and Technological Trajectories: A Suggested Interpretation of the Determinants and Directions of Technical Change," *Research Policy*, 11, 147-162.

DOW, S. P., J. FORTUNA, D. SCHWARTZ, B. ALTRINGER, D. L. SCHWARTZ, and S. R. KLEMMER (2012): "Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results," *Design Thinking Research*, 47-70.

ERIKSSON, T., A. POULSEN, and M.-C. VILLEVAL (2008): "Feedback and Incentives: Experimental Evidence," IZA Discussion Paper No. 3440.

ERAT, S., and V. KRISHNAN (2011): "Managing Delegated Search Over Design Spaces," *Management Science*, forthcoming.

FAGERBERG, J. (2006): *The Oxford Handbook of Innovation*, Oxford University Press.

FREEMAN, R. B., and A. M. GELBER (2009): "Prize Structure and Information in Tournaments: Experimental Evidence," NBER.

FULLERTON, R. L., and R. P. McAFEE (1999): "Auctioning Entry into Tournaments," *American Economic Review*, 107, 573-605.

GATIGNON, H., M. L. TUSHMAN, W. SMITH, and P. ANDERSON (2002): "A Structural Approach to Assessing Innovation: Construct Development of Innovation Locus, Type, and Characteristics," *Management Science*, 48, 1103-1122.

GIROTRA, K., C. TERWIESCH, and K. T. ULRICH (2010): "Idea Generation and the Quality of the Best Idea," *Management Science*, 56, 591-605.

GRADSTEIN, M., and K. A. KONRAD (1999): "Orchestrating Rent Seeking Contests," *The Economic Journal*, 109, 536–545.

GREEN, S. G., M. B. GAVIN, and L. AIMAN-SMITH (1995): "Assessing a Multidimensional Measure of Radical Technological Innovation," *IEEE Transactions on Engineering Management*, 42, 203-214.

GREENHALGH, C., and M. ROGERS (2006): "The Value of Innovation: The Interaction of Competition, R&D and IP," *Research Policy*, 35, 562-580.

HILBE, J. M. (2011): *Negative Binomial Regression*, Cambridge University Press, 2011.

HIGHSMITH, J. A. (1978): "Solving Design Problems More Effectively," *MIS Quarterly*, 2.4, 23-30.

HOLMSTROM, B. (1982): "Moral Hazard in Teams," *Bell Journal of Economics*, 13, 324-340.

JEPPESEN, L. B., and K. R. LAKHANI (2010): "Marginality and Problem-Solving Effectiveness in Broadcast Search," *Organization Science*, 21, 1016-1033.

KATZ, R. (1982): "The Effects of Group Longevity on Project Communication and Performance," *Administrative Science Quarterly*, 27, 81-104.

KELLER, R. T. (1986): "Predictors of the Performance of Project Groups in R & D Organizations," *Academy of Management Journal*, 29, 715-726.

KLINE, S. J., and N. ROSENBERG (1986): "An Overview of Innovation," The Positive Sum Strategy, National Academy Press, ed. R. Landau, Washington: National Academy Press, 275-305.

KLUGER, A. N., and A. DENISI (1996): "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory," *Psychological Bulletin*, 119, 254–284.

KORNISH, L. and K. ULRICH (2011): "Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas," *Management Science*, 57, 107-128.

KORNISH, L. and K. ULRICH (2012): "The Importance of the Raw Idea in Innovation: Testing the Sow's Ear Hypothesis," working paper.

KUZNETS, S. (1962): "Inventive Activity: Problems of Definition and Measurement," *The Rate and Direction of Inventive Activity: Economic and Social Factors*, NBER, 19-52.

LAKHANI, K. R., L. B. JEPPESEN, P. A. LOHSE, and J. A. PANETTA (2007): "The Value of Openness in Scientific Problem Solving," Harvard Business School Working Paper, 07-050.

LAZEAR, E. P., and S. ROSEN (1981): "Rank-order Tournaments as Optimum Labor Contracts," *Journal of Political Economics*, 89, 841-864.

LEE, B., S. SRIVASTAVA, R. KUMAR, R. BRAFMAN, and S. R. KLEMMER (2010): "Designing with Interactive Example Galleries," Proceedings of the conference on Human Factors in Computing Systems, ACM, 2257-2266.

LI, Y., and W. VANHAVERBEKE (2009): "The Effects of Inter-Industry and Country Difference in Supplier Relationships on Pioneering Innovations," *Technovation*, 29, 843-858.

MARCH, J. (1991): "Exploration and Exploitation in Organizational Learning," *Organization Science*, 2: 71-87.

MARSH, R. L., J. D. LANDAU, and J. L. HICKS (1996): "How Examples May (and May Not) Constrain Creativity," *Memory & Cognition*, 24, 669-680.

MERTON, R. K. (1942): *The Sociology of Science: Theoretical and Empirical Investigation*, Chicago: University of Chicago Press.

METCALFE, J.S. (1994): "The Economics of Evolution and the Economics of Technology Policy," *Economic Journal*, 104, 931−944.

MORGAN, J., and R. WANG (2010): "Tournaments for Ideas," *California Management Review*, 52: 77-97.

MULKAY, M. (1975): "Three Models of Scientific Development," *Sociological Review*, 23, 509−526.

NELSON, R. R., and S. WINTER (1982): *An Evolutionary Theory of Economic Change*, Cambridge, MA: Harvard University Press.

PILLER, F. T., and D. WALCHER (2006): "Toolkits for Idea Competitions: A Novel Method to Integrate Users in New Product Development," *R&D Management*, 36, 307-318.

POETZ, M. K., and M. SCHREIER (2012): "The Value of Crowdsourcing: Can Users Really Compete with Professionals in Generating New Product Ideas?" *Journal of Product Innovation Management*, forthcoming.

SCOTCHMER, S. (2004): *Innovation and Incentives*, The MIT Press.

SESHADRI, S., and Z. SHAPIRA (2003): "The Flow of Ideas and Timing of Evaluation as Determinants of Knowledge Creation," *Industrial and Corporate Change*, 12, 1099-1124.

SHALLEY, C. E., and G. R. OLDHAM (1997): "Competition and Creative Performance: Effects of Competitor Presence and Visibility," *Creative Research Journal*, 10, 337-345.

SKINNER, B. F. (1948): "Superstition in the Pigeon," *Journal of Experimental Psychology*, 38, 168-172.

SMITH, S. (2003): "The Constraining Effects of Initial Ideas," *Group Creativity: Innovation Through Collaboration*, Oxford University Press, 15-31.

SMITH, S. M., T. B. WARD, and J. S. SCHUMACHER (1993): "Constraining Effects of Examples in a Creative Generation Task," *Memory & Cognition*, 21, 837-845.

SOUKHOROUKOVA, A., M. SPANN, and B. SKIERA. (2007): "Creating and Evaluating New Product Ideas with Idea Markets," working paper.

TAYLOR, C. R. (1995): "Digging for Golden Carrots: An Analysis of Research Tournaments," *American Economic Review*, 85, 872-890.

TERWIESCH, C., and K. T. ULRICH (2009): *Innovation Tournaments: Creating and Selecting Exceptional Opportunities*, Harvard Business School Press.

TERWIESCH, C., and Y. XU (2008): "Innovation Contests, Open Innovation, and Multiagent Problem Solving," *Management Science*, 54, 1529-1543.

TOUBIA, O., and L. FLORÈS (2007): "Adaptive Idea Screening Using Consumers," *Marketing Science*, 26, 342-360.

TJOSVOLD, D., and L. T. MCNEELY (1988): "Innovation through Communication in an Educational Bureaucracy," *Communication Research*, 15, 568-581.

VAN DE VEN, A. H. (1986): "Central Problems in the Management of Innovation," *Management Science*, 32, 590-607.

VON HIPPEL, E. (2005): *Democratizing Innovation*, Cambridge, MA: MIT Press.

VOSNIADOU, S. (1994): "Capturing and Modeling the Process of Conceptual Change," *Learning and Instruction*, 4, 45-69.

WALTER, T., and A. BACK (2011): "Towards Measuring Crowdsourcing Success: An Empirical Study on Effects of External Factors in Online Idea Contest," 6th Mediterranean Conference on Information Systems.

WARD, T. B. (1994): “Structured Imagination: The Role of Category Structure in Exemplar Generation,” *Cognitive Psychology*, 27, 1-40.

WOOTEN, J., and K. ULRICH (2012): “Idea Generation and the Role of Feedback: Evidence from Field Experiments with Innovation Tournaments,” working paper.

WOOTEN, J., and K. ULRICH (2013): “The Impact of Visibility in Innovation Tournaments: Evidence from Field Experiments,” working paper.

YANG, J., L. A. ADAMIC, and M. S. PAVLOU (2008): “Crowdsourcing and Knowledge Sharing: Strategic User Behavior on TaskCN,” 9th ACM conference.

YANG, Y., P. CHEN, and P. PAVLOU (2010): “Managing Open Innovation Contests in Online Market,” working paper.